BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES

Universidad de Murcia

Propuesta y desarrollo de un modelo para la evaluación de la recuperación de información en internet

Francisco Javier Martínez Méndez

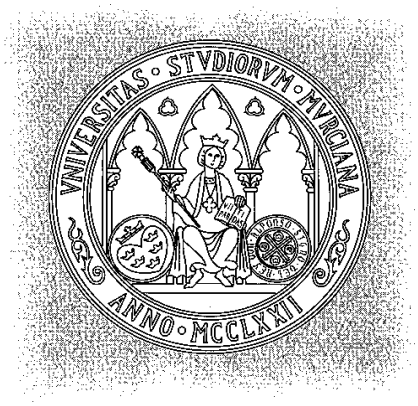
Tesis de Doctorado

Facultad: Ciencias de la Documentación

Director: Dr. José Vicente Rodríguez Muñoz

2002

UNIVERSIDAD DE MURCIA



Facultad de Ciencias de la Documentación

Departamento de Información y Documentación

TESIS DOCTORAL

PROPUESTA Y DESARROLLO DE UN MODELO PARA LA EVALUACIÓN DE LA RECUPERACIÓN DE INFORMACIÓN EN INTERNET

FRANCISCO JAVIER MARTÍNEZ MÉNDEZ

Director: Dr. D. José Vicente Rodríguez Muñoz

2002

"Si buscas algo distinto, no hagas siempre lo mismo"

Albert Einstein

BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES

El esfuerzo y dedicación necesarios para la realización de esta tesis doctoral, es sólo una humilde y pequeña contribución al desarrollo de nuestra Universidad, de la cual nunca dejaré de sentirme orgulloso de formar parte.

La alegría y emoción de mi familia y del director de este trabajo ante su conclusión, es fiel reflejo del cariño y devoción que merecen.

Finalmente, un recuerdo a todas las personas que me quieren, a quienes seguro he abandonado demasiado tiempo y en algún caso decepcionado. Aunque a veces puedan dudarlo, siempre han estado, están y estarán conmigo.

Gracias por vuestro amor y cariño.

Murcia, mayo de 2002

SUMARIO DE CONTENIDOS

Introducción i

Capítulo 1: El método de trabajo 1

Capítulo 2: La Recuperación y los Sistemas de Recuperación de Información

Hacia una definición de la Recuperación de Información. 11

Sistemas de Recuperación de Información. 17

Vista funcional de un SRI. 17
Evolución de los SRI. 19

Modelos para la recuperación de información. 20

La recuperación de información en la web. 23

Breve perspectiva histórica de la web. 23
Métodos de recuperación de información en la web. 25

Los motores de búsqueda como paradigma de la recuperación de información en Internet. 34

Funcionamiento de un motor de búsqueda. 35
Arquitectura de un motor de búsqueda. 36
Los índices de los motores. 42
Tipos de robots. 44
Funcionamiento de los robots. 45
Indización de las páginas. 46
Alineado de los documentos (ranking). 49

Confianza en el funcionamiento de los motores de búsqueda. 51

Tablas e Ilustraciones. 54

Capítulo 3: Evaluación de los SRI.

Necesidad de la evaluación de los SRI. 57

***Relevancia* vs *Pertinencia*. 60**

Las primeras evaluaciones. 64

Proyectos Cranfield. 65

MEDLARS. 67

SMART. 67

El proyecto STAIRS. 69

Conferencias TREC. 70

Medidas tradicionalmente empleadas. 71

Medidas basadas en la *relevancia*. 74

Medidas orientadas al usuario. 79

Cálculo de la *Precisión* y de la *Exhaustividad*. 81

Medidas Promedio E-P. 86

Medidas alternativas a E-P o medidas de valor simple. 88

Modelo de Swets. 91

Modelo de Robertson. 93

Modelo de Cooper. 93

Exhaustividad y *Precisión* normalizadas. 95

Ratio de deslizamiento. 98

Satisfacción y Frustración. 99

Medida de Voiskunskii. 101

Ejemplo práctico de cálculo de medidas de valor simple. 104

Tablas e Ilustraciones. 105

BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES

Capítulo 4: La evaluación de los SRI en la web.

Introducción. 107

Estudios explícitos. 109

Aspectos formales y prestaciones en la recuperación de información. 110
Tamaño del índice del motor de búsqueda. 111
Audiencia del motor. 117
Número de consultas realizadas. 119
Tráfico redirigido. 120
Porción de la página indexada. 121
Fidelidad. 122

Sinopsis de estudios explícitos. 122

Estudios experimentales. 124

Revisión de estudios realizados. 124

Chu y Rosenthal (1996). 124
Leighton y Srivastava (1995-1999). 126
Arents (1995-2001). 127
Bharat y Broder (1997). 128
Ralph (1997). 128
Wishard (1998). 129
Lawrence y Giles (1998). 129
Gordon y Pathak (1999). 130
Gwizdka, J. and Chignell (1999). 134
Ming (2000). 136
Notess (2000). 137
Ljoslund (1999-2000). 138
CNET search site olympics (Thomas, 2002). 143

Sinopsis de estudios experimentales. 143

Propuestas globales de evaluación. 146

Oppenheim (2000). 147
Savoy y Picard (2001). 151
Schlichting y Nilsen (1997). 154
Johnson, Griffiths y Hartley (2001). 156

Tablas e ilustraciones. 163

Capítulo 5: Propuesta del modelo de evaluación.

Introducción. 167

Planteamiento de un Procedimiento de Evaluación. 171

Selección de la muestra objeto de estudio. 171

Determinación del tipo de búsqueda. 172

Establecimiento de criterios para analizar las páginas web. 172

Verificación de la pertinencia de los documentos. 172

Agrupamiento de los índices. 174

Objeto de la evaluación. 174

Desarrollo del Procedimiento de Evaluación. 175

Fase 1: Selección de la muestra objeto de estudio. 175

Fase 2: Elección del tipo de búsqueda y formulación de la pregunta. 181

Fase 3: Establecimiento de criterios previos al análisis de los resultados. 184

Fase 4: Análisis de los resultados. 191

Fase 5: Normalización de la precisión y de la exhaustividad. 191

Fase 6: Determinación de la similitud. 192

Fase 7: Agrupamiento de los índices. 199

Ejemplo de evaluación. 199

Tablas e ilustraciones. 207

Capítulo 6: Aplicación de la propuesta de modelo de evaluación de los SRI en la web.

Elección del conjunto de motores. 211

Temática de las búsquedas. 215

Generalidad de los motores de búsqueda..... 215
Relación de cuestiones realizadas..... 215

Datos relacionados con la *Precisión*..... 219

Precisión de las operaciones con intersección. 219
Precisión de las operaciones con intersección y frase literal. 224
Comparativa de las medias de los valores obtenidos. 227

Datos relacionados con la *Exhaustividad* 227

Exhaustividad de las operaciones con intersección. 228
Exhaustividad de las operaciones con intersección y frase literal. 230

Datos sobre total de documentos localizados. 231

Datos sobre enlaces erróneos y duplicados..... 233

Datos sobre enlaces erróneos..... 233
Datos sobre enlaces duplicados. 236

Datos sobre la similitud de los resultados. 239

Medias de las similitudes obtenidas. 239
Medias de las similitudes obtenidas motor a motor. 240
Similitud de resultados obtenidos por los motores con ambas modalidades de búsqueda en el mismo tema. 243

Datos sobre la medida de valor simple de Borko. 244

Datos sobre la medida de valor simple de Borko tipificada. 247

Datos sobre los agrupamientos de los motores..... 250

Agrupamientos con muestras de diez documentos. 251
Agrupamientos con muestras de veinte documentos. 252
Agrupamientos con muestras de treinta documentos..... 254

Reflexión general sobre el comportamiento de los motores de búsqueda analizados.256

Tablas e ilustraciones. 258

Capítulo 7: Conclusiones 262

Capítulo 8: Referencias 273

BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES

0 Introducción

BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES

Introducción.

La efectividad de la Recuperación de Información.

Inmersos como nos encontramos en un contexto tan tecnológico y supuestamente avanzado como es el de la denominada *Sociedad de la Información y del Conocimiento*, surgen ante nosotros una serie de incógnitas en la apreciación de la efectividad de la recuperación de la información, muy similares a los que pudieron plantearse en tiempos remotos los encargados de la mítica (y felizmente reconstruida) *Biblioteca de Alejandría,* espacio donde *Tolomeo I*, antiguo general de caballería de *Alejandro Magno,* quiso recopilar todo el conocimiento universal o a los que pudo proponerse igualmente, muchos años más cerca de nuestros días, el científico norteamericano *Vannevar Bush*, al plantear su imaginario sistema *MEMEX,* destinado igualmente a almacenar todo el conocimiento universal y facilitar su difusión.

Este conjunto de dudas, que parece acompañar tradicionalmente a la recuperación de información, emana del desconocimiento que el usuario posee sobre una cuestión muy específica: el usuario no sabe si el sistema le proporciona la respuesta más pertinente y eficaz con respecto a sus necesidades de información, o si por el contrario, el sistema le devuelve sólo una parte sesgada de esa respuesta ideal debido a su inadecuado funcionamiento, en definitiva, si satisface o no sus necesidades de información de forma suficiente y necesaria.

Esta situación de incertidumbre ha propiciado que, de forma paralela al desarrollo de los primeros sistemas de recuperación de información, surgiera un conjunto de trabajos y metodologías de evaluación destinadas a ponderar la efectividad de los mismos, convirtiéndose en una de las líneas de investigación más destacadas dentro del campo de la recuperación de información.

Si bien los primeros sistemas de recuperación de información son contemporáneos a los iniciales desarrollos de la red *Internet*, realmente han de transcurrir casi veinte años para que la extensión global de la red y la explosión de la información que ha producido la creación de la *World Wide Web* (mucho más importante y pronunciada que la acaecida a finales de la *II Guerra Mundial*), suscite la implantación y puesta en funcionamiento de los sistemas más grandes y voluminosos que se han conocido: *los motores de búsqueda en Internet*.

A pesar del considerable avance, y plasmación de un amplio conjunto de postulados teóricos, que estos nuevos sistemas constituyen para la investigación en recuperación de la información, incorporando además progresos procedentes de la *Inteligencia Artificial*, de la *Indización Automática* y del *Procesamiento del Lenguaje Natural*, las amplias

dimensiones de las colecciones de documentos que se manipulan se prestan a reproducir estas incógnitas y a acrecentar los niveles de incertidumbre existentes sobre la efectividad de su funcionamiento.

La evolución de los sistemas de recuperación de información.

Una vez superada la etapa en la que se confundía la recuperación de datos con la recuperación de información y delimitadas claramente las diferencias existentes entre ambos conceptos, comienzan a desarrollarse una serie de modelos lógicos para el diseño de sistemas capaces de gestionar las cada vez más ingentes colecciones de datos.

Todos estos modelos coinciden en un planteamiento inicial, largamente defendido por el investigador británico *Thomas Daniel Wilson*, los sistemas deben ser eficaces y dinámicos para apoyar la toma de decisiones de los gestores de las organizaciones.

Conscientes del cada vez mayor número de documentos de los que se dispone, los sistemas deben superar la clásica tendencia del modelo relacional de datos de proporcionar toda la información coincidente con una serie de parámetros de búsqueda, abrumando al usuario con una ingente cantidad de documentos. En esta situación, el usuario difícilmente podrá distinguir aquellos que posean mayor relevancia con respecto a su necesidad informativa, ya que el examen del ingente conjunto de documentos devueltos le resultará, en la práctica, un problema mayor que el no disponer de ellos.

Los sistemas de recuperación de información han evolucionado en la línea de presentar al usuario su respuesta ordenada con base en unos criterios de afinidad a su necesidad informativa, expresada ésta en términos de una expresión regular de búsqueda donde se plasman los conceptos que más le interesan.

Este alineamiento de la respuesta se convierte en la piedra angular del adecuado funcionamiento de un sistema de recuperación de información, en tanto que, cuanto mejor sea el rendimiento del algoritmo que cada uno implementa, mejor calidad tendrá esa respuesta.

Otro aspecto donde se ha evolucionado sustancialmente, no menos importante que el anteriormente indicado, es el desarrollo de la interface de usuario y de las posibilidades de los lenguajes de recuperación de información implementados en estos sistemas. De hecho, estos dos aspectos son de los más utilizados por gran parte de la comunidad científica a la hora de desarrollar procedimientos de evaluación de la efectividad de este tipo de sistemas.

La necesidad de la evaluación de los sistemas de recuperación de información.

Albert Einstein, en su discurso de inauguración del curso de la *Academia Prusiana de Ciencias* de 1914, bromeaba sobre los físicos teóricos al decir que estos científicos podrían realizar muchas cosas pero nunca las que uno quiere que haga en un momento determinado. Esta reflexión del eminente científico alemán puede trasladarse perfectamente a los sistemas de recuperación de información, ya que los mismos son capaces de devolvernos abundante información tras una búsqueda, que igualmente puede resultarnos muy poco útil porque ninguno de esos documentos sea el preciso para satisfacer nuestras necesidades.

La evaluación de estos sistemas nos permitirá identificar cuáles de ellos superan, en mayor o menor medida, el problema anterior, ajustando su respuesta a las necesidades de los usuarios.

Existe un cierto contraste en el tratamiento que al tema de la evaluación de la efectividad de la recuperación de información le confieren los principales manuales especializados. En algunos de ellos este tema se aborda de forma muy concisa y en otros se presenta de forma suficiente. En lo que sí coinciden casi todos los textos consultados es en citar la realización por parte de *Cyril Cleverdon* de los primitivos tests de *Cranfield* y los interesantes trabajos realizados sobre la base de datos MEDLARS por *Frederick W. Lancaster*.

Los análisis de la efectividad de estos sistemas se han concentrado en cuatro grandes grupos: los centrados en la interface del sistema y en las posibilidades de recuperación de información, los centrados en analizar a los distintos sistemas desarrollados comercialmente y, en tercer lugar, los basados en la medida de la relevancia de los documentos. El último grupo lo forman trabajos donde se intentan desarrollar una serie de medidas que evitan hacer uso del concepto de relevancia de los documentos, ante la considerable dosis de subjetividad que implícitamente acompaña a este concepto.

Aunque para nosotros resulten de mayor interés los trabajos del tercer y cuarto grupo, a cuya exposición se dedica el comienzo de esta tesis doctoral, las aportaciones realizadas por los trabajos del primer y segundo grupo no resultan en absoluto superficiales, siendo muchas de ellas citadas en dicha revisión.

Los parámetros más empleados en la determinación basada en la relevancia de la efectividad de los sistemas de recuperación de información, son la *precisión* y la *exhaustividad* de las operaciones de búsqueda, ratios que se pueden calcular por varios métodos, de los que destaca el desarrollado por *Gerard Salton* [SAL, 1983] de pares de valores

"exhaustividad relativa – precisión relativa", que sirve de base para el desarrollo de nuestro experimento.

Como ninguna de estas dos medidas por sí solas ofrece una visión global del funcionamiento de estos sistemas, se han desarrollado diversas medidas de valor simple, a partir de ellas, con la idea de poder establecer una clasificación de la efectividad de una operación de recuperación de información basada en un valor ordinal simple, no en un par de valores. Varias son las propuestas de medida de valor simple elaboradas por diversos autores: *Borko*, *Heine*, *Meadow*, *Vickery* [MEA, 1993] y *Voiskunskii* [VOI, 1997], todas ellas definidas a partir de los pares de valores obtenidos para la *exhaustividad* y la *precisión*.

Igualmente, se han desarrollado otra serie de medidas algo más complejas: modelos de *Swets*, *Robertson* y *Cooper* [RIJ, 1999], ratio de deslizamiento y medidas de satisfacción y frustración [KOR, 1997]. Este conjunto de medidas evita hacer uso de las consideraciones sobre la relevancia de los documentos, por considerarlo un proceso subjetivo.

Se ha podido verificar a lo largo del trabajo que las medidas de valor simple, por regla general y salvo en situaciones excepcionales, adoptan todas ellas un comportamiento equivalente, aunque algunas se ajustan mejor que otras a esos casos específicos, que también se comentan y que no se han presentado en modo alguno en la realización de nuestros experimentos. Este es el motivo por el que optamos por elegir la medida de valor simple I_1 definida por *Borko*, por resultar la de cálculo más simple e intuitivo (es el resultado de la suma de los valores de *exhaustividad* y de *precisión*).

La evaluación de los SRI en la web.

Quizá nunca estuvo en el ánimo de *Tim Berners-Lee* ni en el de *Paul F. Kunz,* cuando desarrollaron hace justamente una década, la *World Wide Web* como mecanismo para el intercambio de la información entre comunidades científicas, el que se elaboraran grandes bases de datos estructuradas que contuvieran todo el contenido de ese nuevo medio de difusión de información. Pero la vertiginosa polularización de la web, fenómeno sin parangón alguno, propició una producción informativa de tal volumen y ritmo de crecimiento que pronto precisó de la puesta en marcha de una serie de mecanismos de clasificación y de localización, de naturaleza diferente a los empleados hasta ese momento.

El primero de estos sistemas desarrollados fue *Yahoo!,* el cual se ha convertido con el tiempo en un inmenso directorio de recursos de información sitos en la web y en uno de los sitios más visitados. Este sistema está gestionado de forma manual por una serie de indizadores que inspeccionan sitios web, los describen y los convierten en un nuevo registro

del directorio. Pero pronto quedó patente que el ritmo de crecimiento de la *World Wide Web* hace imposible un seguimiento humano, precisando que esta serie de procesos se realice de forma automática. Es el momento del desarrollo de los primeros *motores de búsqueda*, sistemas de recuperación de información que poseen una doble vertiente, por un lado una serie de rutinas informáticas se encargan de la recopilación de referencias de sitios web para su base de datos y del otro, una interface de búsqueda permite a sus usuarios intentar localizar la información deseada.

La necesidad crítica de evaluación que poseen los sistemas de recuperación de información, propicia que las evaluaciones de estos nuevos sistemas surjan de forma casi simultánea a su puesta en marcha, y que paralelamente, comience la comunidad de usuarios a plantearse dudas sobre cuál de ellos es el que mejor responde a sus necesidades y cuál de ellos posee la mayor porción de documentos de la web.

El conjunto de evaluaciones desarrolladas sobre la efectividad de los motores de búsqueda no es uniforme. Un primer grupo de estudios se ha centrado en las características de la interface gráfica de usuario y en las posibilidades de recuperación de información. Un segundo grupo se ha dedicado a medir esta efectividad con base en la relevancia de los documentos de la respuesta de cada motor. Y un tercer grupo de estudios, mucho más disperso, se dedica a medir, entre otros aspectos: el ritmo de crecimiento de los motores, el número de referencias incorrectas que poseen en sus índices, su popularidad y el grado de solapamiento existente entre ellos.

Los trabajos de los dos últimos grupos son bastante más costosos de elaborar y, debido a ello, su número es algo inferior a los englobados en el primer grupo. Sus resultados resultan, evidentemente, dispares según los criterios seguidos en su desarrollo, el conjunto de motores que han sido analizados y el período de tiempo en que se han realizado.

Reclamación de una perspectiva de evaluación más próxima al contexto de la web.

La *World Wide Web* presenta una serie de características tan peculiares, que la aplicación de las medidas de evaluación de los sistemas de recuperación de información en este contexto presenta una cierta distancia con la realidad.

Al igual que en la evaluación de los sistemas tradicionales se desarrollaron medidas orientadas al usuario como alternativa a las medidas basadas en la relevancia, consideramos adecuado plantear una propuesta de evaluación que incluya una serie de medidas orientadas no sólo hacia el usuario, *sino también al contexto* en el cual el usuario trabaja, la *World Wide Web*,

entorno muy influenciado por las amplias dimensiones de las colecciones de documentos manejadas, por la volatilidad de los documentos (que conlleva, consecuentemente, obsolescencia en los índices de los sistemas de recuperación de información), y el problema de las insuficientes (algunas veces, inexistentes) descripciones de los documentos realizadas de forma automática.

Esta serie de medidas orientadas a la web no tienen por qué ser necesariamente nuevas, sino que pueden ser fruto de un proceso de adaptación de algunas medidas utilizadas en evaluaciones de los sistemas tradicionales, aunque se debería conferir, sin duda alguna, una mayor importancia a aquellas más vinculadas con la web.

Llegados a este punto de la cuestión, surge de nuevo un problema, ya citado anteriormente: el alineamiento de los documentos en la respuesta. Este proceso cobra particular importancia en este caso concreto, ya que el sistema de recuperación de información en la web seguramente devolverá un extensísimo conjunto de referencias de respuesta, que el usuario no va a tener tiempo (ni quizá interés) en consultar más allá de las veinte o treinta primeras referencias. Por ello, resulta especialmente necesario para estos sistemas, garantizar que los documentos que ocupen los primeros lugares en la respuesta sean los más relacionados con la necesidad de información planteada.

Un aspecto que convierte a este punto en especialmente interesante, no es otro que el ocultismo del funcionamiento del algoritmo de alineamiento del que hacen gala estos sistemas, ya que sus diseñadores apenas han adelantando algunas líneas básicas sobre el mismo, con la única excepción del motor *Google* que sí ha hecho público su algoritmo de alineamiento.

Los últimos trabajos analizados en esta tesis doctoral apuntan a que la evaluación de este tipo de sistemas debe abordarse desde una perspectiva muldimensional, visión que contemple no sólo los aspectos de la evaluación basados en la relevancia de las operaciones de búsqueda, sino también otra serie de aspectos relacionados con la calidad de la respuesta y con el funcionamiento global del sistema.

Es por ello que se va a proponer un modelo para la evaluación de los sistemas de recuperación de información en la web, diseñado bajo esta perspectiva multidimensional a la que se hace referencia anteriormente, propuesta que, una vez demostrada su viabilidad, permita a los usuarios de estos sistemas discernir adecuadamente las posibles abundancias y las carencias de los mismos.

El conjunto inicial de medidas empleadas se extrae de la amplia revisión bibliográfica realizada de forma previa al planteamiento de la propuesta, tanto de las evaluaciones de los sistemas tradicionales como de los desarrollados de manera específica para la web, aunque a la mayor parte

de ellas se les han incorporado unas matizaciones que les permiten aproximarse aún más al contexto donde se va a desarrollar la propuesta de evaluación.

Tras el planteamiento de la propuesta de evaluación, metodológicamente se ha procedido a la realización de un experimento empírico conducente a analizar, de forma inductiva, el comportamiento general de estos sistemas, con la idea de establecer un comportamiento común entre ellos.

Los resultados obtenidos en este proceso de experimentación nos han permitido identificar un amplio porcentaje de aspectos (tanto positivos como negativos) del funcionamiento de este tipo de sistemas de recuperación de información, verificando, al mismo tiempo la viabilidad de nuestra propuesta de modelo para su evaluación e identificando algunas posibles mejoras a aplicar, tanto al propio modelo de evaluación como a los sistemas de recuperación de información.

BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES

1 El método de trabajo.

RESUMEN: El esbozo de una propuesta de modelo de evaluación del funcionamiento de unos sistemas, debe concebirse bajo la óptica de un método inductivo, que permita extraer conclusiones generales sobre ese funcionamiento a partir de la observación del resultado de una amplia serie de experimentos. Este conjunto de ensayos se ha llevado a cabo sobre seis motores de búsqueda, elegidos a partir de un conjunto mayor al cual se le aplicó un proceso de filtrado. El total de observaciones realizadas para analizar el funcionamiento de estos sistemas alcanza la cifra de 10.800 consultas y se han analizados dos de las modalidades más empleadas en la recuperación de información: las búsquedas basadas en la intersección de los términos y la combinación de esta operación con la búsqueda por frase literal.

El método de trabajo.

Necesidad de un método empírico.....2

Nuestra hipótesis.....4

Desarrollo de la hipótesis.....4

Metodología seguida.....6

Necesidad de un método empírico.

La realización de una propuesta de análisis multidimensional del comportamiento de los SRI en la web, que pretenda identificar las posibles analogías y diferencias existentes entre ellos con el objeto de representar un posible comportamiento tipo, se encuentra estrechamente vinculada a la corriente filosófica del *inductivismo*, empleada en la investigación dentro del campo de las *Ciencias Empíricas*. Desde este posicionamiento, el conocimiento científico se concibe como un conocimiento verdadero o cierto sobre la estructura del mundo derivable por inducción a partir de la observación de un cierto número de hechos particulares, tal cual va a ser el caso de nuestro experimento.

Para Fernández, "en lo esencial, dicho método consiste en la derivación -mediante un proceso de inferencia no-deductiva- de un conjunto de reglas generales bajo las cuales queden comprendidos las instancias particulares del fenómeno objeto de estudio.

Estas reglas vendrían a describir una regularidad empírica en el comportamiento de los fenómenos observados, por lo que suele denominárseles 'leyes empíricas'" [FER, 2001]. Estas leyes constituyen, para el investigador inductivista, el núcleo de la ciencia.

Uno de los problemas fundamentales que se presentan, en experimentaciones diseñadas bajo esta perspectiva metodológica, es que la argumentación que sirva de base para la formulación de las leyes empíricas no constituya un razonamiento lógicamente válido. Esta circunstancia propiciaría, en principio, que la conclusión a la que se llegara podría ser falsa, aún cuando las premisas iniciales –los hechos observados- sean verdaderas.

Para superar este problema, "los inductivistas se vieron obligados a apelar a algún principio que permitiera legitimar la pretensión de verdad de las leyes empíricas" [FER, 2001]. El mismo es conocido como "principio de inducción", donde se considera que si en una gran variedad de circunstancias se

observa un gran número de objetos de una cierta clase y todos poseen la propiedad *j*, entonces puede aceptarse a todos los efectos como verdadero que todos los objetos de esa clase poseen la propiedad *j* en cuestión.

Si algún científico encarna en su figura y obra esta corriente de pensamiento, no es otro que el inglés *Isaac Newton*, quien, desde una perspectiva casi *aristotélica,* afirma que "la investigación de los problemas difíciles por medio del análisis debe ir precedida siempre por el método de la composición. El análisis consiste en hacer experimentos y observaciones, y en derivar a partir de ellos conclusiones generales por inducción, rechazando todas las objeciones excepto las basadas en experimentos u otras formas de conocimiento seguro" [COL, 1990].

Y aunque Newton reconoce que la argumentación por inducción, a partir de experimentos y observaciones, no alcanza siempre la demostración de las conclusiones generales, "sigue siendo el argumento más compatible con la naturaleza de las cosas y puede contemplarse como el más fuerte, en la medida en que la inducción sea más generalizada Y si no ocurren excepciones en los fenómenos, la conclusión puede aceptarse como general Pero si en cualquier momento posterior ocurre alguna excepción en los experimentos, entonces debe enunciarse incluyendo las excepciones conocidas" [COL, 1990].

No obstante, sobre este método subyacen unas ciertas dosis de incertidumbre, ya que aún debidamente cumplimentado, no garantiza que una cierta ley general no pueda ser refutada, habida cuenta que es lógicamente posible encontrar un contraejemplo que la invalide, por lo que el inductivismo primitivo reconsideró el carácter de verdad perenne de sus leyes empíricas, adoptando en su lugar, el concepto de verdad probable.

Trasladando esta serie de conceptos generales a nuestra línea de trabajo, la identificación de un comportamiento patrón de los distintos sistemas a la hora de la evaluación de una serie de parámetros, implicaría libertad de elección para el usuario por cualquiera de estos motores.

En cambio, si alguno de los motores presentara una línea de comportamiento diferente, de mejor o peor nivel, el usuario tomaría nota de la conveniencia o no de su uso, en términos de una elevada dosis de certeza, sobre la que se basaría una serie de recomendaciones, aunque nunca podrían considerarse en términos de verdad absoluta, más si cabe en un contexto tan heterogéneo y volátil como es la web.

Nuestra hipótesis.

La hipótesis de partida de nuestro estudio se basa en la elaboración de una propuesta de desarrollo de *un modelo de evaluación de la efectividad de la recuperación de información efectuada por los motores de búsqueda en Internet*, diseñada bajo un enfoque multidimensional. Esta visión multidimensional consta de cinco planos:

1. Cálculo de una medida de valor simple basada en la exhaustividad y en la precisión adaptadas al contexto de la World Wide Web.
2. Tipificación de esta medida de valor simple con respecto al volumen de la respuesta que se le entrega al usuario.
3. Determinación de los porcentajes de enlaces duplicados y de la frecuencia de actualización de los índices de los motores a partir del porcentaje de errores en las referencias ofrecidas.
4. Establecimiento de la similitud existente entre los índices de los motores en función del contenido y del alineamiento de su respuesta.
5. Identificación de posibles agrupamientos entre los distintos motores con base en su similitud.

La elaboración de esta propuesta tiene que ir necesariamente acompañada del desarrollo de una experimentación destinada a verificar su viabilidad por medio del contraste de los resultados obtenidos en nuestro estudio con los aportados en trabajos anteriores realizados por otros autores. Este necesario proceso de contraste se realizará con los estudios más destacados en cada uno de los apartados de nuestra propuesta, cuando ello sea posible, en tanto que nuestro enfoque multidimensional va a incluir aspectos hasta ahora escasamente analizados.

Desarrollo de la hipótesis.

En primer lugar, nuestra propuesta va a hacer uso de una medida de valor simple (como en el caso de la medida I_1 de *Borko* [MEA, 1993]), que incorporará adicionalmente una serie de considerandos concernientes con el alineamiento de los documentos en la respuesta, ya que se va juzgar como más relevante al documento pertinente ubicado entre los primeros lugares de la respuesta que al que aparezca treinta o cuarenta posiciones más atrás, en tanto que la tendencia generalizada de los usuarios no es llegar tan lejos en su consulta.

De este modo, esta medida de valor simple incorporará dos facetas suplementarias: una orientación al contexto de la web y una orientación al

usuario, ofreciendo así un resultado mucho más integral que el asignado tradicionalmente en otros procesos de evaluación.

Para establecer los grados de relevancia de un documento se definirá una *función discreta de valoración de la relevancia*, diseñada bajo una óptica dual: la lógica binaria *documento relevante-documento no relevante* y el ordinal del documento en la respuesta.

Nuestra medida I_1 *adaptada* incorporará además otra serie adicional de elementos de aproximación al contexto de los sistemas evaluados, ya que se establecerán una serie de criterios destinados a analizar las situaciones propias que se presentan en la web y que no suelen presentarse en otros sistemas de recuperación de información, tales como: redireccionamientos a otras páginas, páginas de similar contenido con distintas direcciones, páginas pertenecientes a portales dinámicos, páginas que ya no existen, páginas duplicadas, etc. De este modo, los valores de exhaustividad y precisión calculados se encontrarán más próximos al contexto que en otros experimentos desarrollados, afectando ineludiblemente a la medida de valor simple I_1, que aumentará su notabilidad en relación a la que podía poseer en las evaluaciones de los sistemas tradicionales.

Un segundo objetivo de nuestra propuesta de evaluación resulta mucho más innovador con respecto a otros trabajos previos. Se va a investigar hasta qué punto los motores de búsqueda presentan mejores resultados de exhaustividad y precisión en relación con el volumen de documentos que ofrecen como respuesta. Sería lógico suponer que un motor que posea un índice mayor de documentos, se encuentra en mejores condiciones de alinear a los documentos relevantes entre los primeros de una respuesta que otro motor que posea un tamaño de índice más reducido. Por lo tanto, los pares de valores de exhaustividad y precisión del primer motor deberían ser mejores que en el caso del segundo, simplemente porque posee más documentos sobre la materia objeto de la pregunta.

Esta hipótesis debe verificarse de alguna manera, y para ello se propone una tipificación de la medida de valor simple I_1 de *Borko* con respecto al tamaño de la respuesta de cada motor, generando así una nueva medida de valor simple que denominaremos I_1 *tipificada*.

Paralelamente al estudio de la exhaustividad y de la precisión, se analizará el número de documentos duplicados y el número de documentos erróneos (principalmente correspondientes a enlaces inactivos) presentes en las respuestas, porcentajes que describirán la calidad de los índices y, especialmente en el caso del segundo, la frecuencia de actualización de los mismos.

En tercer lugar se va a analizar la similitud existente entre los índices de los motores de búsqueda analizados. Para ello se adaptarán los resultados de las búsquedas realizadas en nuestro experimento a un espacio vectorial donde se aplicará una función de determinación de la similitud, la del *coseno* más concretamente.

Anteriores experimentos desarrollados sobre sistemas de recuperación de información tradicionales y otros sistemas de la web, han calculado el nivel de solapamiento existente, medida de cierto parecido a la similitud, pero que ofrece menos conclusiones que la aportada por este estudio, porque además de determinar el nivel de coincidencia existente entre los documentos de las respuestas de dos motores, nuestra medida de similitud analizará el comportamiento que ambos tienen a la hora de llevar a cabo su alineamiento, ya que se obtendrán mayores valores de similitud si los documentos coincidentes poseen igual valor de relevancia según la *función discreta de valoración de la relevancia*, al ser la función de coseno es directamente proporcional al producto escalar de los vectores de resultados.

Este cálculo de la similitud de los motores de búsqueda nos ofrecerá una información muy válida sobre la posibilidad de establecer una serie de agrupamientos basados en la distancia existente entre los mismos, procurando detectar aquellos que pudieran resultar más similares y, por tanto, cuyos resultados sean más o menos afines. La determinación de estos posibles agrupamientos constituirá la última fase de esta propuesta de evaluación multidimensional y contribuirá a clarificar muchas dudas de los usuarios sobre la utilidad de algunos motores de búsqueda.

Metodología seguida.

No se puede abordar el proceso de elaboración de una propuesta multidimensional de la evaluación de los sistemas de recuperación de información en Internet si previamente no se llevan a cabo una serie de tareas previas, tales como son la conceptualización de la recuperación de información y de los sistemas que la realizan, junto con un análisis exhaustivo de toda la serie de medidas propuestas para la evaluación de este tipo de sistemas en trabajos precedentes, tanto de los sistemas tradicionales como de los sistemas diseñados para la web.

Las *fuentes de información* empleadas en este estudio se pueden clasificar en dos grandes grupos, según el tipo de sistema analizado. Para abordar el estudio de la evaluación de los sistemas de recuperación de información tradicionales, hemos consultado diversos manuales de referencia obligada para su consulta en este campo, tales como las obras (en mayor medida que otros también incluidos en el repertorio de fuentes) de *Rijsbergen* [RIJ,

1999], *Blair* [BLA, 1990], *Salton* [SAL, 1983] y *Korfhage* [1997]. El estudio de la evaluación de los sistemas de recuperación de información en la web se ha nutrido especialmente de diversos artículos científicos publicados en las principales revistas del campo de las *Ciencias de la Información*, junto a una exhaustiva recopilación de trabajos y estadísticas publicados en la propia *World Wide Web* y en los propios sitios especializados en la evaluación de estos sistemas.

Copia de gran parte de estas referencias electrónicas se han ido depositando en un sitio web de carácter repositorio que se ha creado a medida que se desarrollaba esta tesis doctoral. Este sitio, que se ha denominado '*Web Search Engines Evaluation: The Repositoy web*', está construido con la idea de ser un punto de referencia en la red Internet sobre esta temática. En la actualidad, este sitio está en la URL: <http://gti1.edu.um.es:8080/javima/> [Consulta: 4 de abril de 2002], estando previsto registrar un dominio de Internet específico para la gestión del mismo.

Tras la realización del análisis de trabajos anteriores a partir de la bibliografía, se ha procedido a seleccionar la muestra objeto de estudio, la cual la forman los motores de búsqueda *Alta Vista*, *All the Web*, *Google*, *Microsoft Network*, *Terra* y *WISEnut*. Esta muestra se ha establecido tras la realización de diversos estudios previos. En primer lugar se ha verificado que el volumen de documentos en idioma Español que poseen estos motores es suficientemente significativo para poder realizar sobre ellos un proceso de evaluación (se han producido casos de evaluaciones desarrolladas por otros investigadores donde algunos motores apenas ofrecían documentos escritos en idioma distinto del Inglés).

La elección de nuestro idioma como el utilizado para la realización de las búsquedas objeto de nuestro experimento no es trivial, sino que se debe a dos razones, la primera es la novedad que representa frente a otros estudios anteriores (casi todos empleaban el idioma Inglés), y en segundo lugar, porque se supone que el mayoritario porcentaje de documentos escritos en idioma Inglés frente a otros idiomas puede camuflar, parcialmente, algunos fallos en el funcionamiento del motor. Un número menor de documentos contribuirá, sin duda alguna, a reflejar mejor las virtudes o defectos de los mismos.

La elección de los seis motores de búsqueda citados anteriormente tampoco es casual, previamente se establecieron una serie de requisitos mínimos que debía cumplir un motor para ser evaluado. Posteriormente se analizó un conjunto de motores y tras una primera discriminación, se comenzó a trabajar sobre un total de nueve, que constituyeron el conjunto inicial del estudio.

La naturaleza dinámica de la *World Wide Web* ha propiciado que, a lo largo de la realización de nuestro experimento, algunos de estos sistemas hayan dejado de funcionar en libre acceso (caso del motor *Northern Light* que se ha convertido en un servicio de valor añadido englobado en la oferta de una empresa de servicios de información), o bien hayan sido absorbidos por otros (*Excite* ha sido absorbido por *Overture*). Otra incidencia acaecida a lo largo del experimento fue la detección del alto grado de semejanza existente entre las colecciones de los motores *Hot Bot*, *All the Web* y *Lycos*, similitud que contribuiría, sin duda alguna, a deteriorar los resultados que se obtuvieran en el estudio. Por tanto, ha sido necesario retirar algunos de estos motores de nuestro estudio para que las condiciones del análisis fueran lo suficientemente equilibradas.

Posteriormente se han llevado a cabo treinta operaciones de búsqueda sobre los seis motores analizados, requiriéndoles información sobre distintos temas en cada una de ellas, empleando la operación de intersección e, igualmente, se han efectuado otras treinta búsquedas sobre los mismos temas empleando de forma combinada la operación de intersección y la búsqueda por frase literal.

Los temas de las preguntas se seleccionaron siendo conscientes de que estos sistemas de recuperación de información no son de uso exclusivo de usuarios expertos en una materia o en técnicas de recuperación de información, sino que igualmente son utilizados para encontrar ofertas de viajes para las vacaciones de verano en las *Islas Canarias*, para saber qué jugador del *Real Madrid* marcó cuatro goles en la final de la Copa de Europa celebrada en *Glasgow* el año 1961 o bien para hallar la dirección de correo postal de la *Presidencia* de la *Ciudad Autónoma de Ceuta*. Esta heterogeneidad de motivaciones para su uso y el criterio de proximidad al usuario que rige nuestro estudio, nos llevó a plantear una serie de temas de carácter generalista, no especializado como se ha hecho en otros estudios anteriores.

Que el número de búsquedas realizadas sea el de treinta no ha respondido a un criterio arbitrario sino a una mera exigencia estadística. Para afirmar sin temor a equivocación alguna que el valor de una media aritmética de una distribución de valores representa fehacientemente a esta distribución, se debe aplicar un contraste paramétrico conocido como la prueba de la *T de Student*, que exige ese número mínimo para su realización. Es por ello que todos los valores medios que se ofrecen como resultado del experimento han sido suficientemente contrastados por este método para asegurar su viabilidad.

A lo largo de la realización del experimento, una vez efectuada cada búsqueda se almacenaba en una base de datos el nombre del motor, la URL del recurso devuelto por este motor, el valor de la función de relevancia de

ese recurso, la pregunta a la que correspondía (con lo cual implícitamente se almacenaba también el tema) y la modalidad de búsqueda utilizada. Con ello se ha creado una base de datos de 10.800 registros en entorno *PHP/MySQL* que ha servido de base para la realización de todos los cálculos posteriores de exhaustividad, precisión, porcentajes de duplicados, porcentaje de documentos erróneos y, por último, el cálculo de las similitudes, a partir de un conjunto de rutinas desarrolladas con este propósito.

Otra característica de esta base de datos es que ha sido gestionada a través de la web con el fin de poder trabajar con ella desde cualquier ordenador conectado en la misma y que las aplicaciones desarrolladas para este proceso de evaluación puedan ser reutilizadas a posteriori en el desarrollo de una web específica de evaluación de los motores de búsquedas por parte de sus diseñadores.

Otras herramientas informáticas que se han empleado han sido el editor de textos *Microsoft Word*, la hoja de cálculo *Microsoft Excel*[1] para el cálculo de los valores de todos los parámetros evaluados en este proceso y del software de análisis estadística *SPSS*[2] para la realización de los contrastes. Todo el conjunto de datos gestionados con estas aplicaciones forman parte del Anexo de esta tesis doctoral.

[1] Microsoft Word y Microsoft Excel son aplicaciones de *Microsoft Corp. Inc*.

[2] Este software es propiedad de *SPSS Inc*.

2 La Recuperación y los Sistemas de Recuperación de Información.

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

La Recuperación y los Sistemas de Recuperación de Información.

Hacia una definición de la Recuperación de Información. 11

Sistemas de Recuperación de Información. 17

Vista funcional de un SRI. 17
Evolución de los SRI 19

Modelos para la recuperación de información. 20

La recuperación de información en la web. 22

Breve perspectiva histórica de la web. 23
Métodos de recuperación de información en la web. 25

Los motores de búsqueda como paradigma de la recuperación de información en Internet. 34

Funcionamiento de un motor de búsqueda. 35
Arquitectura de un motor de búsqueda 35
Los índices de los motores. 42
Tipos de robots 44
Funcionamiento de los robots. 44
Indización de las páginas. 46
Alineado de los documentos (ranking). 48

Confianza en el funcionamiento de los motores de búsqueda. 50

Tablas e Ilustraciones. 53

Hacia una definición de la Recuperación de Información.

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

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

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

Otro conjunto de autores expresan las diferencias que, a su juicio, presentan ambos conceptos (con lo cual la definición de recuperación de información queda, en cierto modo, supeditada a la anterior), un tercer grupo de autores la define de forma muy genérica sin entrar en mayores consideraciones sobre estas diferencias, y un cuarto y último grupo pasa de largo sobre este problema, profundizando más en la explicación de los sistemas de recuperación de información[3] (SRI en adelante).

El primer grupo de definiciones debe su naturaleza a la clara influencia de la tecnología informática, cuya evolución ha inducido a muchos autores a cometer el error de considerar sinónimos ambos conceptos, llegándose a olvidar que se puede recuperar información sin procedimientos informáticos (aunque no es lo más común hoy en día, evidentemente), pero el hecho del frecuente y necesario empleo de una tecnología no debe sustituir el adecuado uso de los conceptos terminológicos. Un claro ejemplo de este desacierto lo encontramos en el *Glosario de la Asociación de Bibliotecarios Americanos*[4], que define el término inglés "information retrieval" como recuperación de la información en primera acepción y como recuperación de datos en una segunda acepción [ALA, 1983], considerando ambos términos sinónimos en Lengua Inglesa[5]. De la misma opinión es el *Diccionario Mac Millan de Tecnología de la Información*, que considera a la recuperación de información como el conjunto de "técnicas empleadas para almacenar y buscar grandes cantidades de datos y ponerlos a disposición de los usuarios" [LON, 1989]

---

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

[4] A.L.A. : American Library Association.

[5] Este Glosario indica que "document retrieval" puede considerarse término sinónimo de "information retrieval".

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

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

De esta misma opinión es Blair, quien dedica gran parte de la presentación de su libro[6] a establecer una clara diferencia entre el término *data retrieval* y el término *information retrieval*, utilizando como criterios distintivos, entre otros [BLA, 1990]:

1. En recuperación de datos se emplean preguntas altamente formalizadas, cuya respuesta es directamente la información deseada. En cambio, en recuperación de información las preguntas resultan difíciles de trasladar a un lenguaje normalizado (aunque existen lenguajes para la recuperación de información, son de naturaleza mucho menos formal que los empleados en los sistemas de bases de datos relacionales, por ejemplo) y la respuesta será un conjunto de documentos que pueden contener, sólo probablemente, lo deseado, con un evidente factor de indeterminación.
2. De lo anterior, se deduce que según la relación entre el requerimiento al sistema y la satisfacción de usuario, la recuperación de datos es

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

determinista y en recuperación de información es *posibilista*, por causa del nivel de incertidumbre presente en la respuesta.

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

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

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

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

Tabla 2.1 Diferencias entre recuperación de datos y recuperación de información. Fuente: Rijsbergen, C.J. *Information Retrieval*. [En línea]. Glasgow, University, 1999. <http://www.dcs.gla.ac.uk/~iain/keith/> [Consulta: 21 de octubre de 2001]

Baeza-Yates expone las diferencias entre ambos tipos de recuperación con argumentos algo menos abstractos que los anteriores, destacando que "los datos se pueden estructurar en tablas, árboles, etc. para recuperar exactamente lo que se quiere, el texto no posee una estructura clara y no resulta fácil crearla" [BAE, 1999].

Para este autor, el problema de la recuperación de información se puede definir como "dada una necesidad de información (consulta + perfil del usuario + ...) y un conjunto de documentos, ordenar los documentos de más a menos relevantes para esa necesidad y presentar un subconjunto de aquellos de mayor *relevancia*" [BAE, 1999]. En la solución de este problema se identifican dos grandes etapas:

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

Baeza-Yates ha venido preocupándose especialmente del tema de las estructuras de datos y de los métodos de acceso a los mismos [BAE, 1992], [BAE, 1999], siendo este autor una verdadera referencia en esta materia.

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

Salton indica que, en principio, no deben existir limitaciones a la naturaleza del objeto informativo y Baeza-Yates incorpora la reflexión siguiente: "la representación y organización debería proveer al usuario un fácil acceso a la información en la que se encuentre interesado. Desafortunadamente, la caracterización de la necesidad informativa de un usuario no es un problema sencillo de resolver" [BAE, 1999].

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

Feather y Storges ven a la recuperación de información como "el conjunto de actividades necesarias para hacer disponible la información a una comunidad de usuarios" [IEI, 1997].

Croft estima que la recuperación de información es "el conjunto de tareas mediante las cuales el usuario localiza y accede a los recursos de información que son pertinentes para la resolución del problema planteado. En estas tareas desempeñan un papel fundamental los lenguajes documentales, las técnicas de resumen, la descripción del objeto documental, etc." [CRO, 1987].

Tramullas Sáez impregna su definición del carácter selectivo comentado anteriormente al afirmar que "el planteamiento de la recuperación de información en su moderno concepto y discusión, hay que buscarlo en la realización de los tests de *Cranfield* y en la bibliografía generada desde ese momento y referida a los mecanismos más adecuados para extraer, de un conjunto de documentos, aquellos que fuesen pertinentes a una necesidad informativa dada" [TRA, 1997].

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

Otro autor de esta tendencia es Korfhage, quien se centra en el almacenamiento y recuperación de la información, considerando a estos procesos como las dos caras de una moneda, aunque no entra a definirlos. Para Korfhage, "un usuario de un sistema de información lo utiliza de dos formas posibles: para almacenar información en anticipación de una futura necesidad, y para encontrar información en respuesta una necesidad" [KOR, 1997]. Realmente este autor se dedica más en presentar las diferencias existentes entre dato, información, señal[9], conocimiento y sabiduría [MEA, 1992], [MAR, 1999].

---

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

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

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

Sistemas de Recuperación de Información.

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

Vista funcional de un SRI.

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

Salton opina que "la recuperación de información se entiende mejor cuando uno recuerda que la información que se procesa consiste en documentos", con el fin de diferenciar a los sistemas encargados de su gestión de otro tipo de sistemas, como los gestores de bases de datos relacionales. Salton entiende que "cualquier SRI puede ser descrito como un conjunto de ítem de información (DOCS), un conjunto de peticiones (REQS) y algún mecanismo (SIMILAR) que determine qué ítem satisfacen las necesidades de información expresadas por el usuario en la petición" [SAL, 1983]

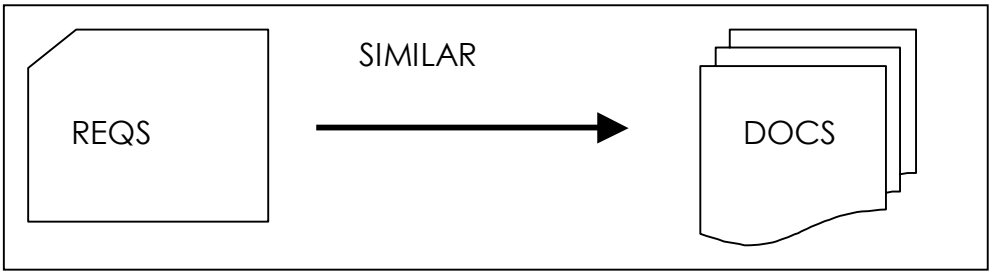


Ilustración 2.1 Esquema simple de un SRI. Fuente Salton , G. and Mc Gill, M.J. Introduction to Modern Information Retrieval. New York: Mc Graw-Hill Computer Series, 1983.

[10] Recogidas en la Tabla 2.1

Salton opina que, en la práctica, este esquema inicial reflejado en la Ilustración 2.1, resulta algo simple y es necesario ampliarlo, "porque los documentos suelen convertirse inicialmente a un formato especial, por medio del uso de una clasificación o de un sistema de indización, que denominaremos LANG" [SAL, 1983]

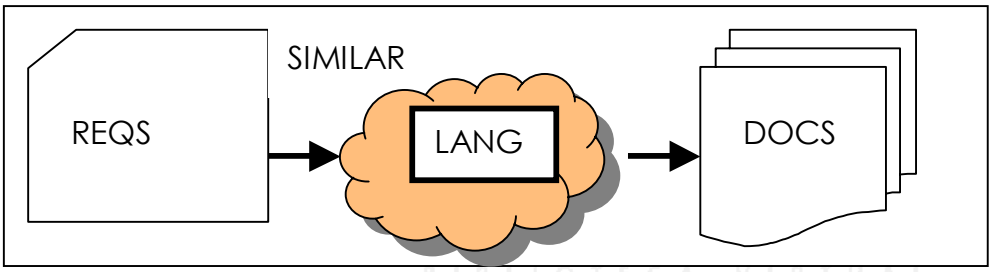


Ilustración 2.2 Esquema avanzado de un SRI. Fuente Salton , G. and Mc Gill, M.J. Introduction to Modern Information Retrieval. New York: Mc Graw-Hill Computer Series, 1983.

En la Ilustración 2.2, se observa que el proceso establecido entre la entrada REQS y SIMILAR es el proceso de formulación de la búsqueda, y el establecido entre SIMILAR y el conjunto de documentos DOCS es el proceso de recuperación. SIMILAR es el proceso de determinación de la similitud existente entre la representación de la pregunta y la representación de los ítems de información. Chowdhury identifica el siguiente conjunto de funciones principales en un SRI [CHO, 1999]:

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

Evolución de los SRI.

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

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

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

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

establecimiento de nuevos términos, tales como WIS ("web information systems") o "sistemas de información basados en la tecnología web destinados a integrarse plenamente con otros sistemas convencionales, llegando a ser más extendidos y de mayor influencia tanto en negocios como en la vida familiar" [WAN, 2001].

Modelos para la recuperación de información.

El diseño de un SRI se realiza bajo un modelo, donde ha de quedar definido "cómo se obtienen las representaciones de los documentos y de la consulta, la estrategia para evaluar la *relevancia* de un documento respecto a una consulta, los métodos para establecer la importancia (orden) de los documentos de salida y los mecanismos que permiten una realimentación por parte del usuario para mejorar la consulta" [VIL, 1997]. Existen varias propuestas de clasificación de los modelos de recuperación, una de las más completas la realiza Dominich, quien establece cinco grupos [DOM, 2000]:

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

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

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

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

Este autor divide a los modelos basados en la recuperación en dos grupos: clásicos y estructurados. En el primero de ellos incluye a los modelos booleano, espacio vectorial y probabilístico.

Posteriormente, presenta una serie de paradigmas alternativos a cada uno de estos modelos: teoría de conjuntos (conjuntos difusos y booleano extendido), algebraicos (vector generalizado, indización por semántica latente y redes neuronales), y por último, probabilísticos (redes de inferencia y redes de conocimiento); los modelos estructurados corresponden a listas de términos sin solapamiento y a nodos próximos (son modelos escasamente difundidos).

Los modelos basados en la navegación entre páginas web son de tres tipos: estructura plana, estructura guiada e hipertexto.

El primero es una simple lectura de un documento aislado del contexto, el segundo incorpora la posibilidad de facilitar la exploración organizando los documentos en una estructura tipo directorio con jerarquía de clases y subclases y el tercero se basa en la idea de un sistema de información que de la posibilidad de adquirir información de forma no estrictamente secuencial sino a través de nodos y enlaces [BAE, 1999].

Es también Baeza-Yates quien proporciona una clasificación adicional de estos modelos de recuperación de información, realizada en función de la modalidad de consulta y de la vista lógica de los documentos:

---

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

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

Tabla 2.3 Clasificación de los Modelos de Recuperación de Información según Baeza-Yates. Fuente: Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999 XX, 513 p.

BIBLIOTECA VIRTUAL

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[14] [GRO, 1998], para conocer las alternativas a los modelos clásicos.

La recuperación de información en la web.

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

---

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

[15] 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.

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

La web[17] es un nuevo contexto, con una serie de particularidades muy definidas, que precisa de una adaptación del concepto de recuperación de información, bajo estas premisas Delgado Domínguez afirma que "se puede definir el objetivo de la recuperación como la identificación de una o más referencias de páginas web que resulten relevantes para satisfacer una necesidad de información" [DEL, 1998]. En este caso, los SRI que se emplean en la web nos van a devolver referencias a los documentos, en lugar de los propios documentos.

Breve perspectiva histórica de la web.

El nacimiento y crecimiento exponencial de la web es un hecho suficientemente conocido y cuyo alcance ha traspasado los límites de la comunidad científica hasta llegar a todo el entorno social. En agosto de 1991, Paul F. Kunz, físico de la *Universidad de Stanford* leyó una noticia en la que difundía la invención de la *World Wide Web* y contactó con Tim Berners-Lee, becario británico del CERN[18]. 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.

---

[17] En el resto de este trabajo se va a utilizar esta manera de denominar al sistema "World Wide Web" para simplificar la lectura.

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

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

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

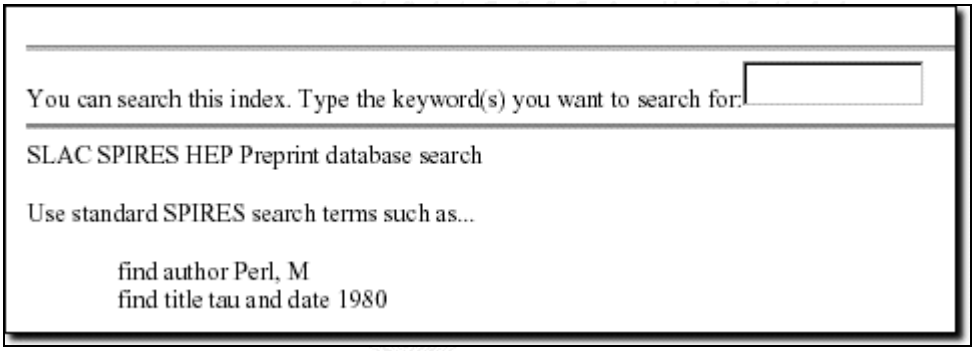


Ilustración 2.3 Sección de la primera página web diseñada por Kunz. Esta página sigue activa en la dirección <http://www.slac.stanford.edu/spires/hep/> de la Universidad de Stanford.

El posterior desarrollo de *Mosaic*, permitió aumentar el número de usuarios de la red Internet que accedía a este novedoso sistema, crecimiento que continuó con el desarrollo de nuevos navegadores: N*etscape* e *Internet Explorer*[19]. A principios del año 1995 se formó la organización *Consorcio World Wide Web*[20] (o *W3C*) que está bajo la dirección del fundador de la Web

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

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

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

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

Métodos de recuperación de información en la web.

Sustancialmente, las técnicas de recuperación de información empleadas en Internet, proceden de las empleadas en los SRI tradicionales, y es por ello que han comenzado a surgir grandes problemas cuando se realizan operaciones de recuperación de información con ellos, en tanto que el entorno de trabajo no es el mismo y las características intrínsecas de los datos almacenados difieren considerablemente. Al mismo tiempo, en la web surgen nuevos problemas, como por ejemplo el peculiar fenómeno denominado *spamming*[21] o los relacionados con el enorme tamaño del índice de estos SRI, que poco a poco llega a alcanzar magnitudes impresionantes, muy difíciles de gestionar adecuadamente con los modelos tradicionales.

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

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

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

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

buscar en sus índices. La tercera, que no está del todo disponible actualmente, es buscar en la web a través de la *explotación de su estructura hipertextual* (de los enlaces de las páginas web[24])" [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[25]), o bien un motor también permite la búsqueda por categorías, como si fuera un directorio (*Microsoft Network*, por ejemplo). Estas situaciones no contribuyen a superar ese grado de confusión.

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

La mayoría de los índices permiten el acceso a los recursos través de dos sistemas: navegación a través de la estructura de las categorías y búsqueda por palabras claves sobre el conjunto de referencias contenidas en el índice. El directorio más grande y famoso es *Yahoo!*, aunque existen otros bastante

---

[24] En algunos textos se emplea el término "hiperenlace" como traducción de "hyperlink".

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

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

Los motores de búsqueda son aplicaciones que manejan también grandes bases de datos de referencias a páginas web recopiladas automáticamente, sin intervención humana. Uno o varios agentes de búsqueda recorren la web, a partir de una lista inicial de direcciones y recopilan nuevas direcciones, generando una serie de etiquetas que permiten su indexación y almacenamiento en la base de datos. Un motor no cuenta con subcategorías como los directorios, sino con avanzados algoritmos de búsqueda que analizan las páginas y proporcionan el resultado más adecuado a una búsqueda. También almacenan direcciones que les son remitidas por los usuarios[26]. Entre los motores más populares destacan *Altavista*, *Lycos*, *Alltheweb*, *Hotbot*, *Overture, Askjeeves, Direct Hit*, *Google*, *Microsoft Network*, *Terra* y *WISEnut,* entre otros. Delgado Domínguez resume en la Tabla 2.4 las características básicas de estos dos métodos de recuperación de información en la web:

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

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

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

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

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

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

---

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

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

desarrollados pueden entenderse como extensiones del lenguaje SQL[29] para el contexto de la web empleado en los tradicionales gestores relacionales.

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

La búsqueda dinámica es "equivalente a la búsqueda secuencial en textos" [BAE, 1999]. La idea es usar una búsqueda online para descubrir información relevante siguiendo los enlaces de las páginas recuperadas. La principal ventaja de este modelo es que se traslada la búsqueda a la propia estructura de la web, no teniendo que realizarse estas operaciones en los documentos que se encuentran almacenados en los índices de un motor de búsqueda. El problema de esta idea es su lentitud, lo que propicia que se aplique sólo en pequeños y dinámicos subconjuntos de la web.

Chang, dentro los SRI en web basados en la recuperación de información por medio de palabras clave, identifica cuatro tipos: motores de búsqueda, directorios, *metabuscadores*[30] y técnicas de *filtrado de información*[31] [CHA, 2001].

Los *metabuscadores* son sistemas desarrollados para mitigar el problema de tener que acceder a varios motores de búsqueda con el fin de recuperar una información más completa sobre un tema, siendo estos mismos sistemas los que se encargan de efectuarlos por el usuario.

Un metabuscador colecciona las respuestas recibidas y las unifica, "la 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

---

[29] SQL: Structured Query Language.

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

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

y crawlers[32]) que tienen registrados y obtienen los resultados que les devuelven. Algunos más sofisticados detectan las URL duplicadas provenientes de varios motores de búsqueda y eliminan la redundancia" [AGU, 2002]. , es decir solo presentan una al usuario.

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

Estos sistemas se diferencian unos de otros en la manera en que llevan a cabo el alineamiento de los resultados[33] en el conjunto unificado[34], y cómo de bien traducen estos sistemas la pregunta formulada por el usuario a los lenguajes específicos de interrogación que maneja cada sistema, ya que el lenguaje común a todos será más o menos reducido.

Algunos metabuscadores se instalan como cliente en entorno local (*Webcompass* o *Copernic*, por ejemplo), o bien se consultan en línea (*Buscopio*, por ejemplo). Otra diferencia sustancial existente entre estos sistemas es la presentación de los resultados, "los llegan a clasificar en dos tipos, los multi buscadores y los meta buscadores: los multi buscadores ejecutan la consulta contra varios motores de forma simultánea y presentan los resultados sin más organización que la derivada de la velocidad de respuesta de cada motor (un ejemplo es *All4One* que busca en una gran cantidad de motores de búsqueda y directorios); los meta buscadores funcionan de manera similar a los multi buscadores pero, a diferencia de éstos, eliminan las referencias duplicadas, agrupan los resultados y generan nuevos valores de *pertinencia* para ordenarlos (algunos ejemplos son *MetaCrawler*, *Cyber411* y *digisearch)*" [AGU, 2002].

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

---

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

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

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

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

Si un metabuscador devolviera todas las referencias de todos los motores y directorios que le sirven de fuente en relación con la materia objeto de una búsqueda, el tiempo de respuesta del sistema alcanzaría valores que seguramente alejarían a los usuarios del metabuscador por excesivo. Es por ello que resulta necesario establecer un número límite de documentos recuperados por motor, con el fin de que el tiempo de respuesta, que de por sí, ya sería siempre mayor que el precisado por un único motor, no aumente excesivamente.

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

Este sistema muestra los resultados progresivamente, es decir a medida que van llegando (tras analizarlos), con lo que el usuario apenas tiene tiempo de espera y las referencias que le entrega el metabuscador son siempre correctas (es decir no va a entregar una dirección de página inexistente o páginas que hubieran cambiado su contenido desde la indización) [BAE, 1999].

Las técnicas de filtrado de la información que comenta Chang son más un complemento de los motores de búsqueda que un modelo alternativo.

El concepto de "filtrado" tiene que ver con la decisión de considerar (a priori) si un documento es relevante o no, eliminándolo del índice en caso contrario.

Estas técnicas "se basan en una combinación de sistemas de autoaprendizaje y sistemas de recuperación de información, y han sido

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

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

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

BIBLIOTECA VIRTUAL

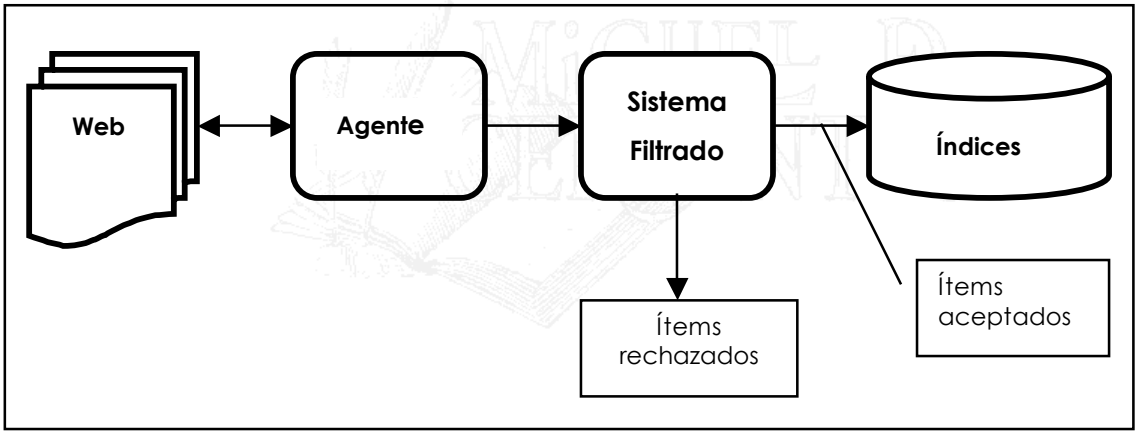


Ilustración 2.4 El proceso de construcción de un motor de búsqueda específico a partir de un filtrado de documentos. Fuente: Chang, G. et al. Mining the World Wide Web: an information search approach". Norwell, Massachusetts: Kluwer Academic Publishers, 2001.

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

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

Hu opina que "los enlaces establecidos entre las páginas web pueden resultar de tremenda utilidad como fuentes de información para los indicadores" [HU, 2001]. El hecho de que el autor de una página web introduzca un enlace en la misma, representa, implícitamente, un respaldo para la página enlazada[35]. La exploración de los enlaces[36] insertados en una página web y, la exploración de los enlaces que apuntan hacia esa misma página, ha propiciado la creación de una nueva familia de motores de búsqueda, de la que *Google*[37] es su mejor exponente [BRI, 1998]. *Google* hace uso de la conectividad de la web para calcular un grado de calidad de cada página, esta graduación se denomina *PageRank* (coincide con el nombre del algoritmo de alineamiento empleado por este motor) y utiliza esta propia capacidad de conexión para mejorar los resultados de búsqueda.

Hu hace referencia, en segundo, tercer y cuarto lugar, a los directorios, motores de búsqueda, metabuscadores y lenguajes de consulta para la web, sistemas que han sido presentados anteriormente.

En quinto lugar menciona la búsqueda de documentos multimedia, campo que se encuentra en plena expansión, máxime cuando cada vez es mayor el número de documentos de esta naturaleza en la web y el número de usuarios que los demandan (seguramente debido a que ha mejorado la capacidad de su conexión a Internet). Para este autor su desarrollo "está considerado uno de los mayores desafíos en el campo de la recuperación de información" [HU, 2001]. El último grupo citado por Hu engloba a los sistemas de recuperación con interface basada en procesamiento de

---

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

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

[37] *Google* es un juego de palabras con el término "googol", acuñado por Milton Sirotta, sobrino del matemático norteamericano Edward Kasner, para referirse al número representado por un 1 seguido de 100 ceros. El uso del término por parte de *Google* refleja la misión de la compañía de organizar la inmensa cantidad de información disponible en la web y en el mundo. Fuente: *Todo acerca de Google* [En línea] Mountain View, CA: Google, 2001. <http://www.google.com/intl/es/profile.html> [Consulta: 21 de enero de 2002].

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

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

De la totalidad de los SRI que se han desarrollado en Internet, los motores de búsqueda son los que más se incardinan con la naturaleza dinámica del contexto de la web, siendo unos sistemas de evolución paralela al crecimiento de la web y al aumento del número de usuarios. Constituyen además uno de los desarrollos más consolidados de las técnicas de *Indización Automática* [SAL, 1983] [GIL, 1999] y, al mismo tiempo, son los sistemas más sensibles a toda la amplia serie de situaciones peculiares que se presentan en la red: "spamming", inaccesiblidad de páginas, deficientes o inexistentes descripciones de las páginas, volatilidad, etc.

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

Si se asumen que de lo completa, representativa y actualizada que sea la colección de un motor de búsqueda, depende su calidad; en un directorio, en cambio, la misma reside en la capacidad de los gestores en la realización de las descripciones y en el número de estos gestores, ambos motivos más relacionados con capacidades presupuestarias que con prestaciones tecnológicas.

En cambio, los motores representan un claro ejemplo de la aplicación de las técnicas de recuperación de información a la resolución de un reto, tan antiguo como moderno, en el campo de la Información y la Documentación: disponer en un índice las referencias a la mayor parte de los documentos existentes.

Funcionamiento de un motor de búsqueda.

El funcionamiento de un motor debe estudiarse desde dos perspectivas complementarias: la recopilación y la recuperación de información. Un motor compila de forma automática las direcciones de las páginas que van a formar parte de su índice tras realizar sobre su contenido un proceso de indización. Una vez se encuentren estos registros debidamente depositados en la base de datos del motor, los usuarios buscarán en su índice por medio de una interface de consulta, que puede ser más o menos avanzada en función del grado de desarrollo del sistema. Al módulo encargado de la recopilación de las páginas se le conoce comúnmente como robot[38], "es un programa que rastrea la estructura hipertextual de la web, recogiendo información sobre las páginas que encuentra. Esa información se indiza y se introduce en una base de datos que será explorada posteriormente utilizando un motor de búsqueda" [DEL, 1998].

Estos robots pueden recopilar varios millones de páginas por día, y actualizar la información recogida en los índices en períodos de tiempo extremadamente pequeños si consideramos la extensión del espacio al que nos estamos refiriendo. Por regla general, se parte de una lista inicial de direcciones de sitios web, que son visitados por el robot, y a partir de ahí cada robot rastrea a su manera la web, de ahí que la información almacenada en cada base de datos de cada motor sea diferente. A diferencia de Delgado Domínguez, Baeza-Yates distingue en un robot las funciones de análisis o rastreo ("crawling") de las de indización o indexación ("indexing"), con lo cual él habla de dos módulos independientes, el "crawler" o robot y el indexador [BAE, 1999].

Arquitectura de un motor de búsqueda.

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

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

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

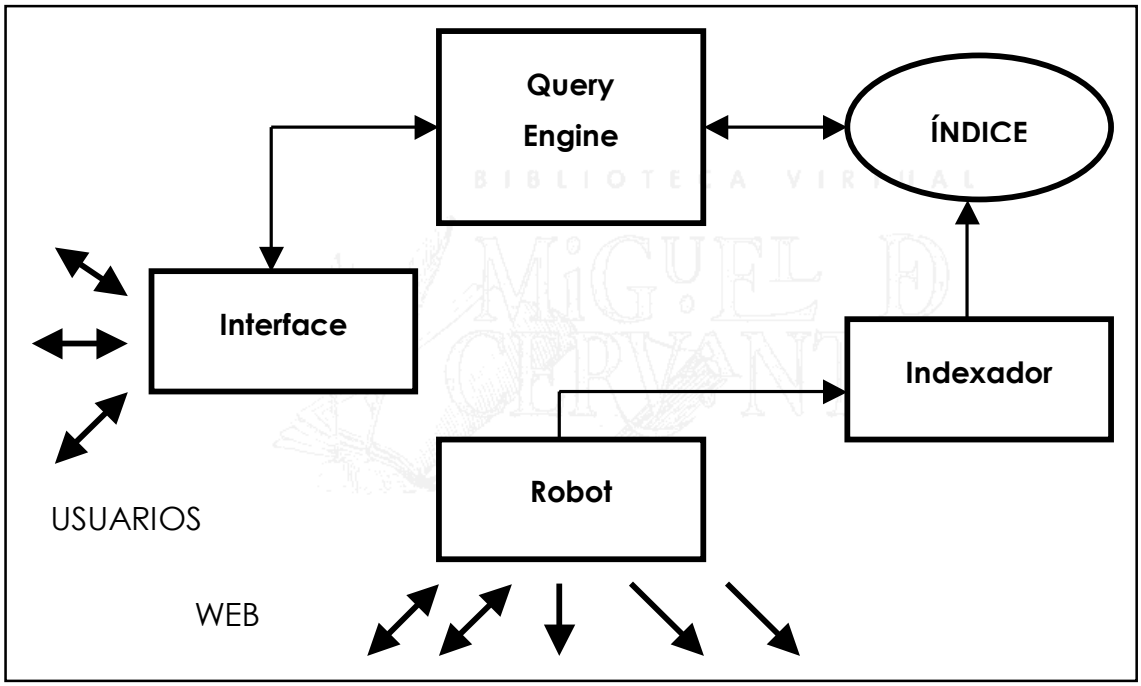


Ilustración 2.5 Arquitectura simple de un motor de búsqueda. Fuente: o a partir de un filtrado de documentos. Fuente Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999 XX, 513 p.

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

- La actualización de los índices es complicada y lenta.

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

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

Estos problemas propician que uno de los campos de estudio más recientes en la web, sea el desarrollo de una serie de alternativas a este modelo de arquitectura simple, para procurar paliar estos defectos. Baeza-Yates destaca la arquitectura del sistema *Harvest*[40] como la más importante de todas. Este sistema es un paquete integrado de herramientas gratuitas para recoger, extraer, organizar, buscar, y duplicar información relevante en Internet desarrollado en la *Universidad de Colorado* [BOW, 1994].

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

[40] *Harvest* significa "cosecha".

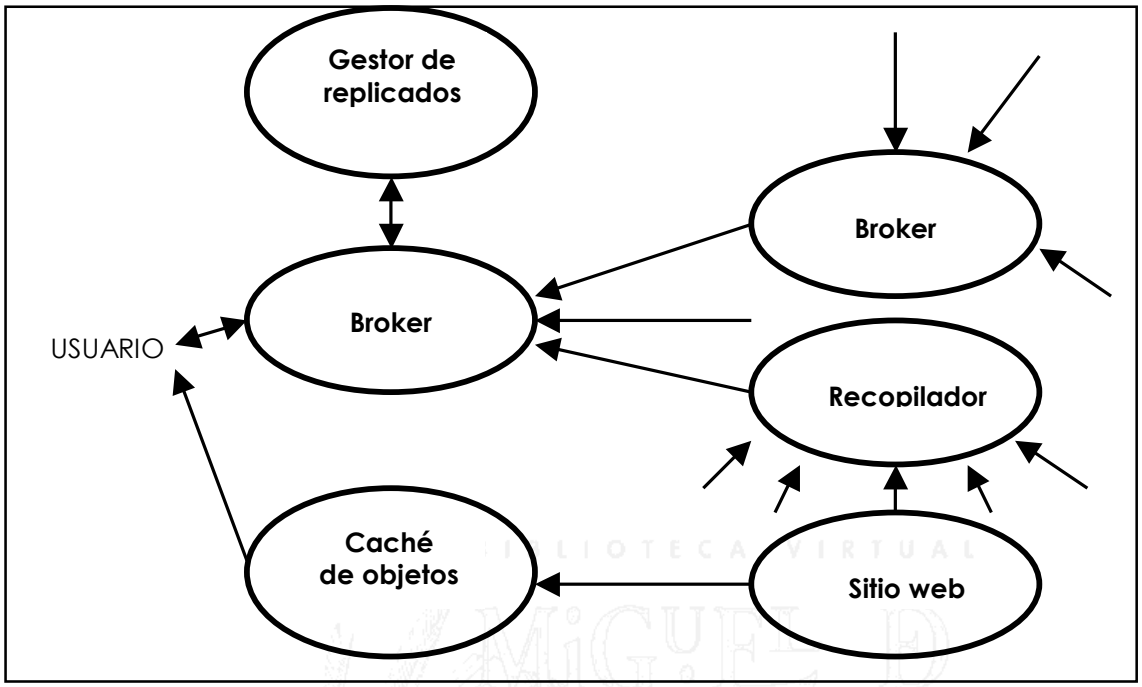


Ilustración 2.6 Arquitectura Harvest. Fuente Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999 XX, 513 p.

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

Esta arquitectura solventa el problema de la carga de tráfico en las conexiones con el motor, porque aumenta la velocidad de conexión con los robots en tanto que estos descartan gran cantidad de contenidos de las páginas que analizan y no las transfieren al entorno local, aliviando considerablemente la carga de tráfico.

En último lugar, la información es recopilada de forma independiente por cada robot, sin tener que realizar una gestión sincronizada.

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

Interface de usuario.

El estudio de la interface debe abordarse bajo dos perspectivas: la interface que el sistema dispone para que el usuario exprese sus necesidades de información (Chang la denomina "interface de consulta" [CHA, 2001]) y la interface de respuesta que dispone el sistema para mostrar al usuario el resultado de su operación de búsqueda [BAE, 1999].

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

Ilustración 2.7 Formulario de búsqueda simple del motor All the Web. Fuente: <http://www.alltheweb.com>

Aunque generalmente los motores de búsqueda suelen disponer de una interface de búsqueda avanzada (como la que se muestra en la Ilustración 2.8), en la cual el usuario puede incorporar a su ecuación de búsqueda una serie de parámetros adicionales, tales como: uso de operadores booleanos, búsqueda por frase literal, búsqueda aplicando operadores de adyacencia, búsqueda por términos opcionales (Baeza-Yates los denomina "invitados", son esos términos que podrían aparecer o no en un documento objeto de una consulta), restricciones geográficas, restricciones por tipo de dominio, restricciones por idioma, etc. Algunos sistemas permiten refinar la búsqueda, es decir, especificar más la pregunta sobre el conjunto de documentos recuperado inicialmente e incluso algunos motores permiten restringir el alcance la operación de búsqueda a alguna de las partes de los

documentos contenidos en sus índices (el caso más común es el título o el texto).

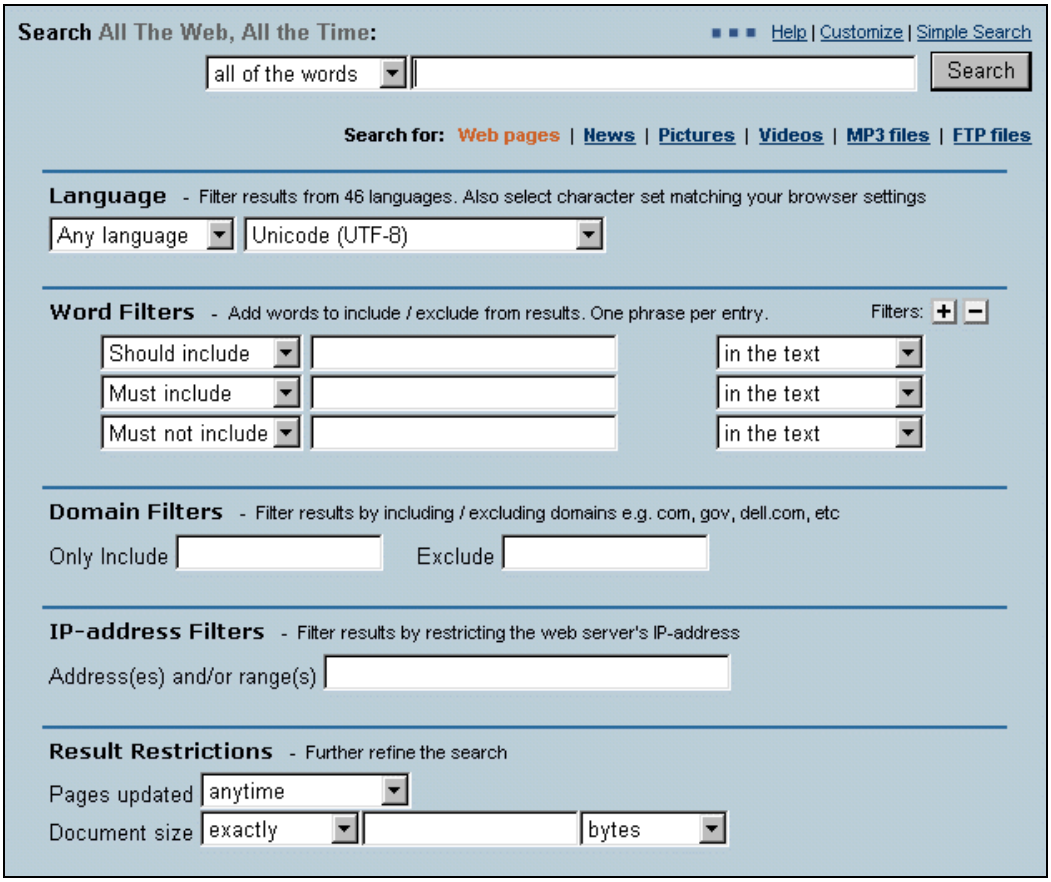


Ilustración 2.8 Sección del formulario de búsqueda avanzada del motor All the Web. Fuente: <http://www.alltheweb.com/advanced>

Son muchos los criterios por los que se pueden identificar diferencias entre las posibilidades de recuperación de información ofrecidas por cada motor. La interface de usuario para la realización de las consultas es una de las más empleadas y constituye uno de los parámetros más empleados en los artículos y páginas web dedicadas a la evaluación de las prestaciones de cada motor [WIN, 1995], [DAV, 1996], [SLO, 1996], [ZOR, 1996] y [WES, 2001].

Otra de las diferencias, algo más interna y no tan explícita, es la forma en la que un sistema interpreta una relación de varios términos como expresión de consulta sin operadores entre ellos (por ejemplo: "Historia Región Murcia"), es decir, cómo descompone el motor la expresión y construye la ecuación de búsqueda. En este punto existen también diferencias entre los sistemas, unos realizan una búsqueda por proximidad de las palabras de la expresión (es el

caso de Overture), otros motores recuperarán documentos donde aparezcan todas las palabras (Google) y otros recuperarán documentos donde al menos aparezca una de las tres palabras de la ecuación (Alta Vista). Chang identifica cinco tipos de búsqueda [CHA, 2001]:

1. Término simple
2. Términos múltiples
3. Basadas en el contexto
4. Lenguaje natural
5. Correspondencia de patrones

La búsqueda por término simple tiene como objeto devolver una colección de documentos donde al menos se pueda encontrar una ocurrencia de ese término, algunos sistemas permiten restringir esa búsqueda a un campo determinado (búsqueda por referencia cualificada). La búsqueda por términos múltiples permite diversas combinaciones basadas en el *Álgebra de Boole*: intersección de los subconjuntos correspondientes a cada término, unión de estos subconjuntos o exclusión de un subconjunto de otro; algunos sistemas permiten la combinación de los operadores para construir expresiones booleanas complejas.

Las búsquedas basadas en el contexto usan los operadores de proximidad, es decir, localizan documentos donde los términos integrantes de la ecuación de búsqueda se encuentren situados en la misma frase o en el mismo campo (además de, por supuesto, el mismo documento). El caso más cercano de proximidad es la adyacencia[41] (cuando los términos están escritos en un orden determinado, por ejemplo, uno a continuación del otro). Algunos motores permiten la búsqueda en lenguaje natural, que puede resultar especialmente interesante para aquellos usuarios no experimentados en el uso de un motor específico o en el empleo de los operadores booleanos o basados en el contexto.

Estos sistemas interpretan cuestiones del estilo de "¿qué jugador de fútbol es el máximo goleador de la Copa de Europa?" o "¿qué ciudad es la capital de Angola?", devolviendo como resultados un conjunto de documentos que han considerado adecuados con la temática de la pregunta efectuada, tras haber sometido a esta expresión a un procedimiento de análisis del

[41] También es conocida por "búsqueda por frase literal" o "búsqueda exacta".

texto, interpretando la necesidad informativa (*Alta Vista* y *Northern Light* implementan esta modalidad). Un caso extremo de procesamiento de las expresiones en lenguaje natural es el motor *Askjeeves,* que llega a simular una "entrevista" con el usuario ya que, tras recibir la cuestión, la interpreta y extrae de su base de conocimientos una serie de cuestiones que traslada al usuario para refinar su exploración en la base de datos y ajustar mejor la respuesta. Por último, algunos sistemas devuelven los documentos por correspondencia con un patrón de caracteres introducido en la interface de consulta. Es el caso de aquellos motores que permiten hacer uso del operador de truncamiento, como es el caso de *Alta Vista*.

Los índices de los motores.

El índice "es el corazón de un motor de búsqueda" [CHA, 2001], generalmente consiste en una lista de palabras con valor de discriminación asociadas a sus correspondientes documentos, que en este caso son las descripciones de los contenidos de las URL recopiladas. La mayor parte de los motores de búsqueda emplean como estructura de datos un *fichero inverso* [BAE, 1992], [TRA, 1997], [RIJ, 1999], [CHA, 2001], [DEL, 2001], basado en la idea general que se muestra en la ilustración siguiente.

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

(a) Example text; each line is one document

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

(b) Inverted file for text of (a)

Ilustración 2.9 Ejemplo de la estructura de un fichero inverso (tabla de la derecha). Fuente: Rijsbergen, C.J. Information Retrieval. [En línea]. Glasgow, University, 1999. <http://www.dcs.gla.ac.uk/~iain/keith/> [Consulta: 21 de octubre de 2001]

En la práctica el fichero inverso se convierte en una enorme estructura de datos con serios problemas de gestión. Los distintos motores de búsqueda se sirven de distintos esquemas para definir estas estructuras de datos. Se cuenta con un parámetro, denominado *granularidad*, que distingue "la

exactitud con la que el índice identifica la localización de una palabra clave; en general, los índices pueden clasificarse con base en este parámetro" [CHA, 2001]. Esta clasificación distingue tres niveles:

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

El índice emplea un conjunto de punteros que apuntan a una tabla donde se recogen todas las URL en las que aparece una palabra clave. La manera en la que se ordenan estos punteros depende de un mecanismo interno de ordenación basado, generalmente, en criterios de frecuencias o pesos en el documento. El enorme tamaño de la colección de URL recopiladas por los motores obliga a buscar formas de simplificar al máximo el tamaño de estos índices. En la Tabla 2.6 se presentan algunas de las diversas técnicas empleadas.

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"), reduciéndose así el número de entradas en el índice
Supresión de las palabras vacías[42]	Se suprimen del índice todas aquellas palabras por las que no tiene sentido recuperar información (artículos, preposiciones, adjetivos o interjecciones, por ejemplo)
Comprensión de textos	Técnicas de compactación del tamaño del fichero

Tabla 2.6 Técnicas empleadas para reducir el tamaño de los índices de un motor de búsqueda. Fuente: Chang, G. et al. Mining the World Wide Web: an information search approach". Norwell, Massachusetts: Kluwer Academic Publishers, 2001.

[42] Las *palabras vacías* son palabras sin valor de discriminación. En Inglés se conocen como "stopwords".

Las tres primeras técnicas consiguen reducir hasta un 30% del tamaño del índice y la cuarta llega, en algunos casos, a comprimir hasta un 90% el tamaño [BAE, 1999]. Resulta clara la tendencia a disminuir el tamaño del índice, ya que cuando las búsquedas constan de varios términos y uno de ellos es muy frecuente, el motor puede tardar varios segundos en responder, hecho no muy bien considerado por muchos autores y profesionales. El uso de índices con *granularidad* más consistente implica menor tamaño de índice y menos punteros, lo que favorece una simplificación de la estructura de datos. Baeza-Yates expone como ejemplo la idea que Glimpse emplea para el sistema *Harvest*: "las preguntas del usuario son resueltas por medio de ficheros inversos, que proporcionan una lista de bloques lógicos que son leídos de forma secuencial, ya que su tamaño es menor" [BAE, 1999].

Tipos de robots.

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

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

Funcionamiento de los robots.

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

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

La extensión de la web genera problemas con el refresco de los índices de los motores, ya que transcurre un necesario período de tiempo entre dos análisis del mismo recurso, intervalo que varía mucho según el motor. Analizando esta problemática, Baeza-Yates esboza una analogía entre el índice y las estrellas del cielo: "lo que vemos en un índice jamás ha existido, ya que la luz ha viajado a lo largo de mucho tiempo hasta llegar a nuestro ojos. Cada página se indexó en un momento distinto del tiempo, pero al ir a ella obtenemos el contenido actual [BAE, 1999].

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

Característica de rastreo	No	Si
Rastreo profundo	Excite	El resto
Soporte de marcos	Excite, FAST	El resto
Mapas de imágenes	Excite, FAST	Alta Vista, Northern Light
Robots.txt	Ninguno	Todos
Metadatos	Excite	El resto
Rastreo por popularidad	Ninguno	Todos
Inclusión pagada	Excite, Google[43]	Alta Vista, Inktomi, FAST

Tabla 2.7 Características de rastreo de los robots de los principales motores de búsqueda. Fuente: Aguilar González, R. Monografía sobre motores de búsqueda [En línea]. Yahoo! Geocities, 2002. <http://www.geocities.com/motoresdebusqueda/crawlers.html> [Consulta: 3 de abril de 2002]

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

Indización de las páginas.

A medida que los robots recopilan páginas, la información contenida en las mismas debe ser indizada, Delgado Domínguez opina que "existen dos estrategias básicas, no mutuamente excluyentes, para realizar este proceso: usar información que provee el creador o editor del documento, o extraerla directamente del documento" [DEL, 1998].

El volumen de información que gestiona un robot obliga a que el motor de búsqueda implemente algún tipo de *indización automática* [GIL, 1999]. En la práctica, los principales motores emplean ambas estrategias para disponer de una completa descripción del contenido de la página analizada. Aguilar González enumera una serie de criterios utilizados para esta descripción: "el título del documento, los metadatos, el número de veces que se repite una palabra en un documento, algoritmos para valorar el peso del documento, etc." [AGU, 2002].

La mayoría de los motores calculan el número de veces que se repiten las palabras claves en el cuerpo de una página, después escudriñan estas palabras en el nombre del dominio o en la URL, posteriormente en el título de la página, en el encabezado y en los metadatos. El orden en que se busca en cada uno de estos elementos varía en función del motor (cada uno usa sus propios algoritmos con criterios diferentes).

Si el motor encuentra las palabras claves en todos estos criterios, entonces posee una razón para asignar un peso mayor al documento. Otra metodología se basa en el número de enlaces que la misma reciba o proporcione.

Aguilar González indica que la primera propuesta en esta línea es de Attardi, de la *Universidad de Pisa*, implementada en el motor *Arianna* y que ha servido de base para el desarrollo de motores que analizan los enlaces (como *Google o WISEnut*) [AGU, 2002].

Un ejemplo representativo del comportamiento de un motor clásico a la hora de indizar las páginas web es el motor *Alta Vista:*

- Da prioridad alta a las palabras del título y a las palabras que están localizadas en el comienzo de la página.
- Asigna mayor peso a una palabra en un documento según su frecuencia absoluta.

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

Google es el mejor ejemplo de uso extensivo de los enlaces como base para mostrar los documentos a los usuarios de un motor. En este motor, la función de indización la llevan a cabo dos módulos: el *indexador* y el *clasificador*. El primero lee las páginas procedentes del *storeserver*[44], descomprime los documentos y selecciona los términos incluidos en los mismos.

Cada documento se convierte en un conjunto de palabras (o '*hits*'), donde se graba la palabra y su posición en el documento, una aproximación de su fuente de texto y otra serie de detalles, por medio del *clasificador*.

El *indexador* analiza también los enlaces incluidos en cada página web, información necesaria para calcular el alineamiento de las páginas a la hora de la recuperación de información [BRI, 1998].

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

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

Características de la 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 2.8 Características de la indización realizada por los principales motores de búsqueda. Fuente: Aguilar González, R. Monografía sobre motores de búsqueda [En línea]. Yahoo! Geocities, 2002. <http://www.geocities.com/motoresdebusqueda/crawlers.html> [Consulta: 3 de abril de 2002]

Alineado de los documentos (ranking).

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

En función del buen funcionamiento de su algoritmo de alineamiento, el motor será mejor o peor valorado por los usuarios del mismo. Si un motor no discrimina su respuesta en función de la *relevancia* con la temática objeto de la pregunta, el usuario encontrará documentos muy relevantes mezclados con otros menos relevantes e incluso con muchos nada relevantes, lo que le obligará a consultar un gran número de los documentos devueltos por el motor, teniendo que visitar muchas pantallas y perdiendo, en consecuencia, un cuantioso tiempo. En esta situación, el usuario terminará por no recurrir a este motor de búsqueda. Si, en cambio, el motor discrimina ese grado de relación, el usuario encontrará entre los primeros documentos a los más relevantes con la temática objeto de la pregunta, por lo que aumentará su grado de satisfacción con el motor y continuará utilizándolo. Tradicionalmente este procedimiento ha sido uno de los secretos

mejor guardados por los responsables de los distintos motores de búsqueda y realmente, no se dispone de una información clara de cómo las motores lo llevan a cabo, con excepción del motor *Google* que ha hecho público su algoritmo *PageRank* [BRI, 1998].

Al igual que ocurría con los criterios de indización existen dos grandes grupos de algoritmos para el alineamiento, los que emplean variantes del modelo de espacio vectorial o del modelo booleano y los que siguen el principio de extensión de los enlaces.

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

El segundo grupo de algoritmos aporta una de las mayores diferencias conceptuales sobre el alineamiento: el uso de los enlaces de cada página (tanto los que recibe una página como los que emanan de ella). El número de enlaces que apuntan a una página sirve como una medida de su popularidad y calidad. La presencia de enlaces comunes entre un conjunto de página es también una medida de relación de los temas tratados en ellas. Dentro de esta nueva tipología de técnicas de alineamiento, identificamos tres clases:

- *WebQuery*: da un alineamiento a las páginas que forman la respuesta a una consulta con base en cómo de conectadas están entre ellas. Adicionalmente, extiende el conjunto de páginas de la respuesta a otra serie de páginas altamente conectadas al grupo original de respuestas.
- *HITS*[45]: 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

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

establece un enlace a una página es porque la considera interesante, y que personas con intereses comunes tienden a referirse a las autoridades sobre un tema dentro de una misma página [ARA, 2000]. Conectores y autoridades son conceptos que se retroalimentan: mejores autoridades son inducidas por enlaces desde buenos conectores y buenos conectores vienen de enlaces desde buenas autoridades [BAE, 1999].

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

Confianza en el funcionamiento de los motores de búsqueda.

Tras analizar el funcionamiento de los motores de búsqueda y conocer las particularidades de los problemas que afectan a la calidad de su funcionamiento, llega el momento de establecer si los mismos son fiables o

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

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

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

Es decir, el problema planteado es explicar razonadamente por qué el motor de búsqueda proporciona unas páginas y no otras, o lo que es lo mismo, se trata de resolver el problema de la asignación de *relevancia* a las páginas devueltas con respecto a la temática de la pregunta planteada. A pesar de las altas dosis de subjetividad que puedan estar presentes en estos postulados anteriores, no dejan de reflejar un problema muy común en el uso de los motores de búsqueda: estos sistemas muchas veces no proporcionan información verdaderamente relevante sobre un tema, a pesar de devolvernos una ingente cantidad de documentos en un tiempo relativamente escaso y de disponer el motor de una enorme base de datos con varios millones de documentos indexados. Este hecho provoca que surjan opiniones tan rotundas como la anterior descalificando por completo la operatoria de estos sistemas.

Partiendo de una postura mucho más positivista, hay que intentar diferenciar los problemas que padecen estos sistemas a la hora de llevar a cabo correctamente su tarea, e intentar aislarlos en su contexto, exponiendo

claramente su alcance y sus posibles soluciones. Esta serie de problemas podrían clasificarse de la siguiente manera:

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

El primer grupo de problemas se encuentra muy ligado, la mayor parte de las veces, a una inadecuada formulación de la ecuación de búsqueda. Este problema, típico en la recuperación de información, cobra más importancia si cabe, en el contexto de los motores de búsqueda cuyos usuarios no tienen por qué disponer de unos conocimientos mínimos en técnicas de recuperación de información. Este problema intenta ser paliado por los responsables de los propios motores quienes, en mayor o menor medida, insertan en la ayuda de estos sistemas explicaciones de cómo sacar el mejor partido al motor para recuperar información. También es frecuente encontrar publicaciones impresas y páginas web que realizan esta labor de asesoramiento al usuario no iniciado. Paralelamente al problema de los legos en la materia, surge el problema de adaptación que sufren algunos usuarios al cambiar de un motor a otro, aunque este problema es de menor incidencia. Con ello, el problema de la formulación inadecuada de las ecuaciones de búsqueda subyace y va a estar, de alguna manera, siempre presente en toda operación de recuperación de información.

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

El tercero de los problemas es el de la inadecuada indización de las páginas web. A la ausencia de una estructura básica de los documentos analizados por los robots, hay que unir lo reciente de esta tecnología (algunos motores con más de cien millones de páginas en su base de datos y aún se consideran "prototipos"). En este punto confluyen muchas circunstancias, "además de por las propias limitaciones de la tecnología en su estado actual, existen también claros intereses, por parte de los propietarios de las páginas web, en que sus páginas aparezcan en la búsqueda y que aparezcan en la mejor posición.

Este interés, legítimo en principio, puede dejar de serlo cuando se utilizan mecanismos que distorsionan la realidad en ese afán por aparecer en los procesos de búsqueda. A modo de ejemplo de esas malas prácticas se puede citar el conocido 'spamming'" [GRA, 2000]. Además de este problema, no hay que olvidar que las técnicas de indización automática ofrecen un rendimiento en absoluto cercano a la perfección, por lo que los algoritmos que implementan los motores cometen fallos que se trasladan al conjunto de resultados.

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

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

Tablas e Ilustraciones.

Tabla 2.1 Diferencias entre recuperación de datos y recuperación de información. Fuente: Rijsbergen, C.J. *Information Retrieval*. [En línea]. Glasgow, University, 1999. <http://www.dcs.gla.ac.uk/~iain/keith/> [Consulta: 21 de octubre de 2001]..... 14

Ilustración 2.1 Esquema simple de un SRI. Fuente Salton , G. and Mc Gill, M.J. Introduction to Modern Information Retrieval. New York: Mc Graw-Hill Computer Series, 1983..... 17

Ilustración 2.2 Esquema avanzado de un SRI. Fuente Salton , G. and Mc Gill, M.J. Introduction to Modern Information Retrieval. New York: Mc Graw-Hill Computer Series, 1983..... 18

Tabla 2.2 Clasificación de los Modelos de Recuperación de Información según Dominich. Fuente: Dominich, S. 'A unified mathematical definition of classical information retrieval'. Journal of the American Society for Information Science, 51 (7), 2000. p. 614-624..... 20

Tabla 2.3 Clasificación de los Modelos de Recuperación de Información según Baeza-Yates. Fuente: Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999 XX, 513 p. 22

Ilustración 2.3 Sección de la primera página web diseñada por Kunz. Esta página sigue activa en la dirección <http://www.slac.stanford.edu/spires/hep/> de la Universidad de Stanford. 24

Tabla 2.4 Características de directorios y motores de búsqueda. Fuente: Delgado Domínguez, A. Mecanismos de recuperación de Información en la WWW [En línea]. Palma de Mallorca, Universitat de les Illes Balears, 1998. <http://dmi.uib.es/people/adelaida/tice/modul6/memfin.pdf> [Consulta: 18 de septiembre de 2001] 27

Ilustración 2.4 El proceso de construcción de un motor de búsqueda específico a partir de un filtrado de documentos. Fuente: Chang, G. et al. Mining the World Wide Web: an information search approach". Norwell, Massachusetts: Kluwer Academic Publishers, 2001. 32

Ilustración 2.5 Arquitectura simple de un motor de búsqueda. Fuente: o a partir de un filtrado de documentos. Fuente Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999 XX, 513 p. 36

Ilustración 2.6 Arquitectura Harvest. Fuente Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999 XX, 513 p..... 38

Ilustración 2.7 Formulario de búsqueda simple del motor All the Web. Fuente: <http://www.alltheweb.com> 39

Ilustración 2.8 Sección del formulario de búsqueda avanzada del motor All the Web. Fuente: <http://www.alltheweb.com/advanced> 40

Ilustración 2.9 Ejemplo de la estructura de un fichero inverso (tabla de la derecha). Fuente: Rijsbergen, C.J. Information Retrieval. [En línea]. Glasgow, University, 1999. <http://www.dcs.gla.ac.uk/~iain/keith/> [Consulta: 21 de octubre de 2001] 42

Tabla 2.5 Clasificación de los ficheros inversos a partir de la *granularidad* de su índice. Fuente: Chang, G. et al. Mining the World Wide Web: an information search approach". Norwell, Massachusetts: Kluwer Academic Publishers, 2001. 43

Tabla 2.6 Técnicas empleadas para reducir el tamaño de los índices de un motor de búsqueda. Fuente: Chang, G. et al. Mining the World Wide Web: an information search approach". Norwell, Massachusetts: Kluwer Academic Publishers, 2001. 43

Tabla 2.7 Características de rastreo de los robots de los principales motores de búsqueda. Fuente: Aguilar González, R. Monografía sobre motores de búsqueda [En línea]. Yahoo! Geocities, 2002. <http://www.geocities.com/motoresdebusqueda/crawlers.html> [Consulta: 3 de abril de 2002] 45

Tabla 2.8 Características de la indización realizada por los principales motores de búsqueda. Fuente: Aguilar González, R. Monografía sobre motores de búsqueda [En línea]. Yahoo! Geocities, 2002. <http://www.geocities.com/motoresdebusqueda/crawlers.html> [Consulta: 3 de abril de 2002] 48

3 Evaluación de los SRI.

RESUMEN: La naturaleza determinista de los SRI propicia su necesidad intrínseca de evaluación. Es por ello que, de forma paralela al desarrollo de su tecnología, ha surgido un amplio campo de trabajo dedicado específicamente al establecimiento de medidas que permitan valorar su efectividad. Un repaso exhaustivo de la bibliografía especializada en esta serie de trabajos 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.

Evaluación de los SRI.

Necesidad de la evaluación de los SRI. 56

***Relevancia* vs *Pertinencia*....59**

Las primeras evaluaciones. 63

Proyectos Cranfield. 64

MEDLARS. 66

SMART. 66

El proyecto STAIRS. 68

Conferencias TREC. 69

Medidas tradicionalmente empleadas. 70

Medidas basadas en la *relevancia*. 73

Medidas orientadas al usuario. 78

Cálculo de la *Precisión* y de la *Exhaustividad*. 79

Medidas Promedio E-P. 84

Medidas alternativas a E-P o medidas de valor simple. 86

Modelo de Swets. 89

Modelo de Robertson. 91

Modelo de Cooper. 91

Exhaustividad y *Precisión* normalizadas. 93

Ratio de deslizamiento. 96

Satisfacción y Frustración. 97

Medida de Voiskunskii. 99

Ejemplo práctico de cálculo de medidas de valor simple. 102

Tablas e Ilustraciones. 103

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].

Pors 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" [POR, 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 consideraba que se llevaban a cabo excesivas evaluaciones sobre determinados aspectos relacionados con el acceso físico, cuando donde deberían realizarse más evaluaciones era 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 van a ser 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 cuestión que, 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 cuestión "¿cuándo un documento es relevante?".

El término *relevancia*, según el *Diccionario Vox de la Lengua Española* [VOX, 2001], es "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].

Para el *Diccionario Vox de la Lengua Española* [VOX, 2001], "*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, 1992].

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" [POR, 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 el trabajo de Sparck-Jones [SPA, 1981], en el cual se recopilan 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*[46], *exhaustividad*[47], tasa de fallo[48], 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;

---

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

[47] 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.

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

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 cuando éste resalta la repercusión de este experimento en los sistemas de indización y en los lenguajes documentales[49] [LOP, 1998]. Aunque donde realmente ha sido realmente importante este experimento es en el campo de la evaluación de los SRI, debido a, principalmente, dos razones: este estudio estableció los factores que más afectan al funcionamiento de los SRI y refrendó el desarrollo de la primera metodología de evaluación de estos sistemas, sobresaliendo la introducción de las medidas de *exhaustividad* y *precisión*, que, bien ligeramente modificadas, bien provistas de nuevas interpretaciones o bien asociadas a exacerbadas críticas, nos siguen acompañando hasta nuestros días.

El segundo proyecto *Cranfield* consistió en 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 información sobre la naturaleza de los fallos de un SRI[50]. Los resultados que proporcionó este nuevo experimento 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.

---

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

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

MEDLARS.

El funcionamiento del sistema MEDLARS[51] 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 realizada estaba orientada a 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%. En este caso, a diferencia del test de *Cranfield*, sí se proporcionaban pormenores sobre las razones de los fallos 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 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 se encaminaban a desarrollar un prototipo de SRI íntegramente automatizado. A lo largo de su progreso, se propusieron medidas añadidas a las planteadas en los tests *Cranfield*, algunas de las cuales se han incorporado a posteriores

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

evaluaciones de SRI[52]. Para la realización del experimento, se establecieron 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 tanto los usuarios que realizan las preguntas como los evaluadores ajenos a esas preguntas están de acuerdo [CHO, 1999].

[52] Se hace referencia a los gráficos *Exhaustividad-Precisión* basados en el promedio de ambos valores y a sus versiones normalizadas. Todas estas series de medidas se presentan posteriormente.

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[53] 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[54], 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 *exhaustividad*es 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

[53]STAIRS: Storage And Information Retrieval System.

[54]Estos bajos valores de *exhaustividad* planteaban una duda en el seno de este contexto, una base de datos legal, que 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.

que las diferencias, tanto de *precisión* como de *exhaustividad* no son muy grandes.

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[55] 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 nace 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[56] 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.

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

[56] DARPA: Defence Advanced Research Projects Agency. Es la Agencia de Proyectos de Investigación del Departamento de Defensa de Estados Unidos de América.

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].

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. Resulta 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 responde citando a Cleverdon, quien en 1966 presentaba seis medidas: "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" [RIJ, 1999]. Este autor opina que las cuatro primeras medidas son intuitivas y fácilmente estimables, y que las dos últimas, 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 recoge las medidas de Cleverdon y cita a Vickery, quien 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)" [CHO, 1999]. El conjunto de medidas de Cleverdon es usado posteriormente por Salton, quien además manifiesta 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 han empleado una amplia serie de medidas basadas en otros criterios. Meadow las sintetiza todas en tres grupos: medidas basadas en la *relevancia*, medidas del proceso y medidas del resultado [MEA, 1992]. Este amplio conjunto de medidas se sintetiza en las siguientes Tablas 3.2:

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 (I) Resumen de las medidas, basadas en la Relevancia de los documentos recuperados, empleadas en la evaluación convencional de la recuperación de la información. Fuente: Meadow, C. T. Text Information retrieval Systems. San Diego: Academic Press, 1993.

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 3.2 (II) Resumen de las medidas, basadas en la evaluación de los procesos, empleadas en la evaluación convencional de la recuperación de la información. Fuente: Meadow, C. T. Text Information retrieval Systems. San Diego: Academic Press, 1993.

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 3.2 (III) Resumen de las medidas, basadas en el Resultado obtenido, empleadas en la evaluación convencional 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 operación de recuperación de información, un usuario recupera un conjunto de documentos, de los cuales unos van a formar parte del subconjunto de 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 otro conjunto de documentos que igualmente serán relevantes con esa temática, y otro conjunto de documentos que no lo van a ser, tal como se muestra en la Ilustración 3.1.

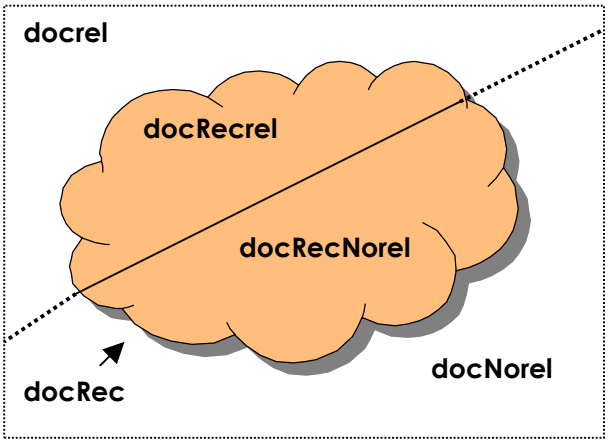


Ilustración 3.1 Distribución resultados operación de búsqueda. docRec: documentos recuperados: docRecrel: documentos recuperados relevantes; docRecNorel: documentos recuperados no relevantes; docrel: documentos relevantes; docNorel: documentos no relevantes.

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 3.3, más conocida como *Tabla de Contingencia*:

	RELEVANT	NON-RELEVANT	
RETRIEVED	$A \cap B$	$\overline{A} \cap B$	B
NOT RETRIEVED	$A \cap \overline{B}$	$\overline{A} \cap \overline{B}$	$\overline{B}$
	A	$\overline{A}$	N

(N = number of documents in the system)

Tabla 3.3 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], nos 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 quedan recogidas 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 3.4 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* refleja el porcentaje de documentos recuperados no relevantes sobre el total de documentos no relevantes de la base de datos. Esta medida cobra especial importancia cuando 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.

En la cita anterior, Salton hace referencia 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 = (E * G) / (E * G) + F (1 – G)			
P = *precisión*	E = *exhaustividad*	G = *generalidad*	F = fallo

Tabla 3.5 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, ya que cuanto mayor es el valor de la *precisión*, menor va a mostrarse el valor de la *exhaustividad*. Esta situación viene recogida en casi todos los manuales que hemos revisado por medio de una gráfica similar a la que reproducimos a continuación del libro de Rijsbergen, donde se representa la evolución de estas dos ratios en dos preguntas diferentes, Q_1 y Q_2 [RIJ, 1999]

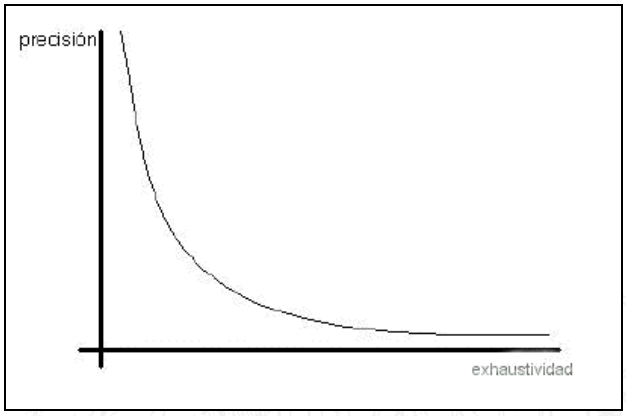


Ilustración 3.2 Evolución típica de la *precisión* y de la *exhaustividad* en un SRI

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.

Esto es, 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 puede ser 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].

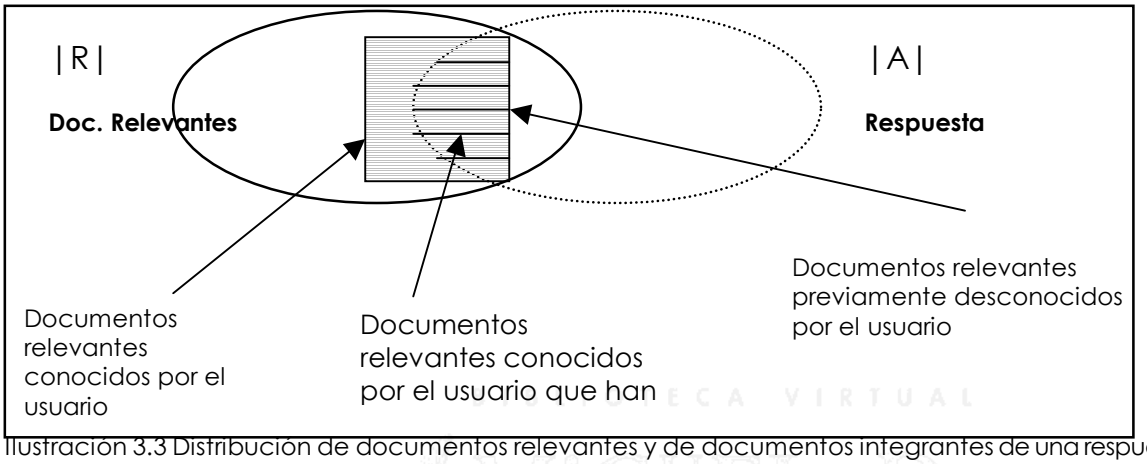


Ilustración 3.3 Distribución de documentos relevantes y de documentos integrantes de una respuesta a una operación de búsqueda. Fuente: Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999.

<u>Cálculo de la *Precisión* y de la *Exhaustividad*.</u>

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, desafortunadamente, 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 una muestra aleatoria de la colección documental, donde el usuario evaluará la *pertinencia* de los mismos y entonces, empleando técnicas estadísticas fiables, estimará el 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. Blair cita a Tague [BLA, 1990] quien avisa de la dificultad de realizar esta tarea en base de datos de muy baja *generalidad* (porcentaje de documentos relevantes 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 incorpora 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 el problema de cómo garantizar una distribución similar 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 o inexistente método de estimación estadística.

Salton [SAL, 1983] apuesta por calcular los valores de *exhaustividad* y *precisión* sobre una muestra de documentos de la colección total. Realmente Salton[57] no llega a afirmar 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 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

[57] Esta postura de Salton la defiende en el curso del estudio de comparación del rendimiento de STAIRS y SMART, que hemos citado anteriormente.

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}. 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 3.6 Cálculo de pares de valores E-P de la búsqueda ejemplo. Fuente: elaboración propia.

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. 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:

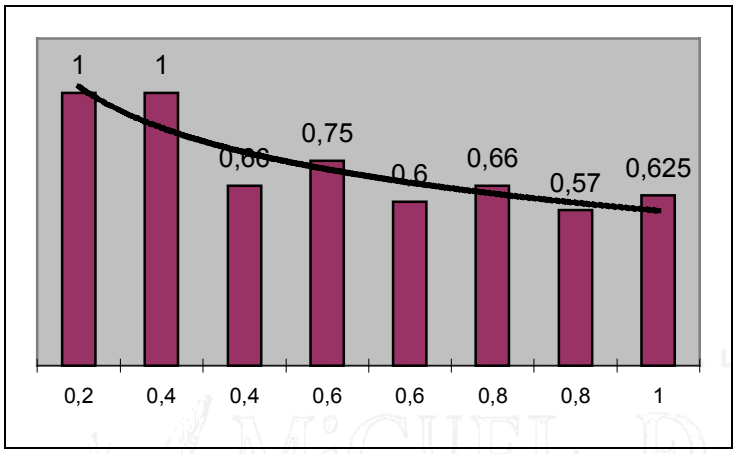


Ilustración 3.4 Evolución distribución pares de valores E-P del ejemplo. Fuente: elaboración propia.

Gráficos como el de la anterior ilustración, han sido fuertemente criticados por no exponer 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 apunta también 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 Q individual, quisiéramos representar los resultados de varias búsquedas más, 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]. Por último, al visualizar 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.

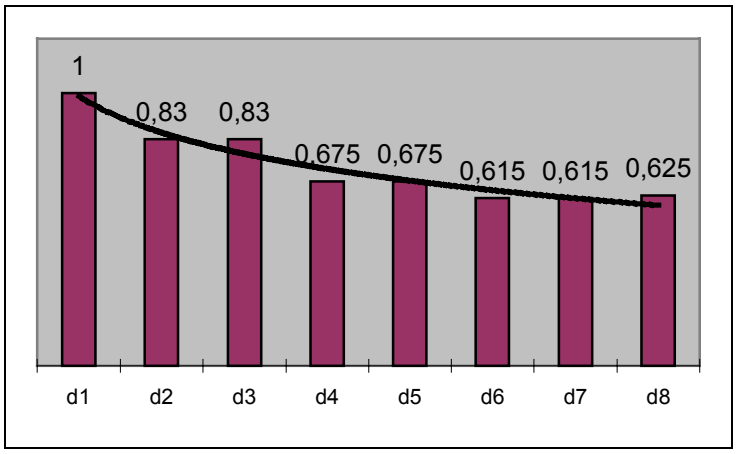


Ilustración 3.5 Evolución distribución valores E-P del ejemplo corregida. Fuente: elaboración propia.

En la anterior Ilustración 3.5 se han determinado 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.

Ante esta serie de problemas, Salton propone el cálculo de 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 3.7 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 nueva formulación permite representar una curva E-P con valores diferentes de *exhaustividad* para cada valor de la *precisión* y, además, la función es continua en lugar de discreta. La representación gráfica de Salton, coincide con la curva propuesta por Rijsbergen [RIJ, 1999]. Esta forma de representar la relación de los pares de valores E-P resulta también válida cuando se realiza una única búsqueda (es decir, el valor de la variable i es

igual a la unidad), sirva como ejemplo la representación de evolución E-P recogida en la Ilustración 3.6, correspondiente a la búsqueda Q del ejemplo:

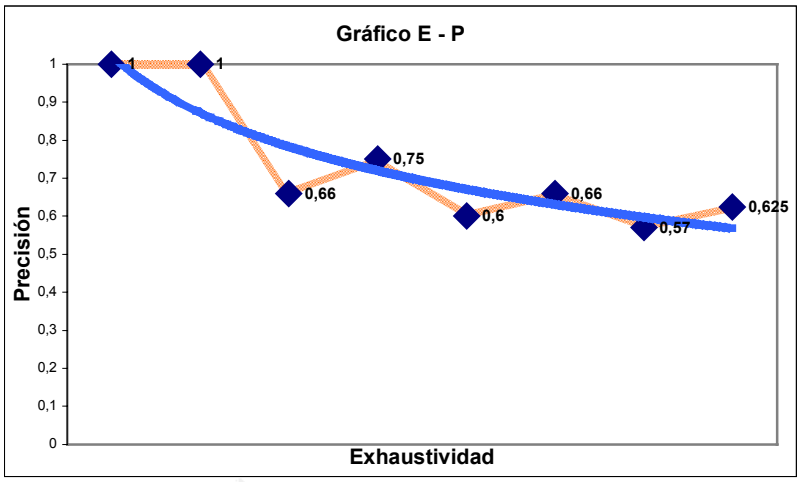


Ilustración 3.6 Representación de la evolución E-P real y óptima del ejemplo. Fuente: elaboración propia.

En esta nueva representación, la curva E-P de la búsqueda Q presenta un recorrido más o menos similar a la presentada por Rijsbergen, que recogen igualmente muchos otros autores.

Este método de cálculo de los valores E-P se conoce también como cálculo de *exhaustividad y precisión relativa,* aunque no existe una unanimidad entre los autores analizados, entendiéndose estas nuevas medias como unas aproximaciones a los verdaderos valores de ambas ratios, sencillas de calcular y que pueden constituir una sólida base para la evaluación de los SRI.

Korfhage presenta dos maneras distintas de calcular el promedio de la *exhaustividad* y de la *precisión*. La primera de ellas consiste en "calcular los promedios de la *precisión* para un conjunto de tres o de once valores previamente establecidos de la *exhaustividad*" [KOR, 1997]. Estos dos métodos son conocidos como "promedio en tres puntos" y "promedio en once puntos".

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 alternativo de medidas de la efectividad de una recuperación de información, medidas todas ellas que pretenden superar los problemas anteriormente analizados de la *exhaustividad* y de la *precisión*. Casi todas hacen uso de técnicas probabilísticas y resultan de cálculo más complicado que las anteriores [RIJ, 1999].

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" [BAE, 1999].

Repasando las presentaciones de los principales autores consultados, Salton denomina a estas medidas como "de valor simple", en tanto que ya no vamos a representar el resultado de una evaluación en función de un par de valores sino de un único y simple valor, que puede ser objeto de clasificación.

Este mismo autor nos presenta, de forma sucesiva, el *Modelo de Swet* (que propicia el desarrollo de 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 nueva faceta de la evaluación de la recuperación de la información, relacionada con las medidas de valor simple: la *satisfacción*.

El mismo autor cita a Borko, para quien este factor es la suma de los valores de *precisión* y de *exhaustividad* obtenidos en una operación de recuperación de información. Otras medidas de esta naturaleza se han definido en diversos trabajos de otros autores y se recogen sintetizadas en la Tabla 3.8.

Medidas de la calidad de una búsqueda

Autor	**Expresión**
Borko	I = E + P
Meadow	M = 1 – SQRT [(1 – P^2) + (1 – R^2)] / SQRT(2) [58]
Heine	D_1 = 1 – 1 / [(1/P) + (1/R) – 1]
Vickery	V = 1 – 1 / [(2/P) + (2/R) – 3]
	Nota: **SQRT** es el símbolo de la raíz cuadrada

Tabla 3.8 Medidas de la calidad de una búsqueda propuestas por varios autores. Fuente: Meadow, C. T. *Text Information retrieval Systems*. San Diego: Academic Press, 1993.

Todo este conjunto de medidas constituye un instrumento útil para valorar la calidad de una búsqueda, aunque, más adelante se verá que diversos autores opinan que no resulta suficiente (en todos los casos) con realizar una simple o compleja operación aritmética, y que hará falta aplicar otras medidas nuevas para estimar esta satisfacción.

Estos autores basan sus opiniones, fundamentalmente, en la escasa capacidad que ellos 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, es justo al contrario, cuanto más cerca se encuentre el resultado obtenido de cero, será mejor la búsqueda.

Para justificar esta afirmación sirve el siguiente ejemplo: un usuario realiza una búsqueda que podría calificarse como "ideal" porque recupera diez documentos de los cuales todos son relevantes y en toda la base de datos sólo se recogen estos diez documentos relevantes.

En este caso "ideal", la búsqueda del usuario tendría un valor de E igual a 1 y un valor de P igual a 1, siendo, por lo tanto, igual a 2 el valor de la medida de Borko (lo máximo que puede alcanzar), mientras que la medida de Meadow tendrá un valor de 1 (su máximo) y las otras dos medidas valdrán 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

[58] Meadow emplea la letra "E" para representar a su medida, pero hemos decidido modificarla con el fin de evitar confusiones con la representación de la *exhaustividad*.

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 valdrá ahora 0.5 y E valdrá 0.25 (en este caso, estamos en un 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.

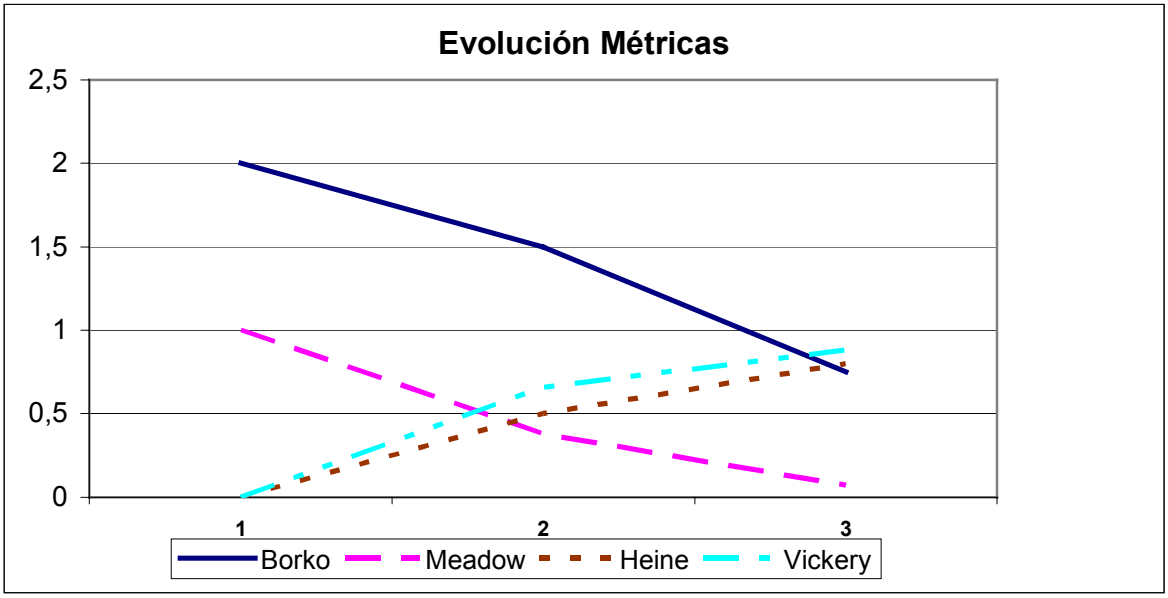


Ilustración 3.7 Representación de la evolución de las medidas de calidad de una búsqueda ejemplo. Fuente: elaboración propia.

Así, cuando más efectiva sea 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. Siguiendo con el repaso de las principales medidas alternativas o de valor simple propuestas por los autores, Rijsbergen nos 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].

En último lugar, Baeza-Yates es mucho más sintético en la presentación de estas medidas, en tanto que para este autor, las *medidas simples* figuran dentro de la determinación de valores de E-P [BAE, 1999]. Posteriormente se refiere a la "Medida E" de Swet (dedicándole poco más de un párrafo) y cita la existencia de otras medidas, sin explicar nada de su cálculo, figurando entre ellas, la medida de *satisfacción* y de *longitud esperada de búsqueda* a las que Korfhage dedica varios apartados.

Lo que sí queda suficientemente claro para la mayoría de los autores es la necesidad de desarrollar un conjunto de medidas alternativas a las tradicionales, tal como ocurre con Salton para quien "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, de forma separada de otra serie de criterios como puede ser el coste.
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 esta serie de premisas se desarrolla el *Modelo de Swets*, suficientemente explicado en [RIJ, 1999] y [SAL, 1983]; este modelo 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 típicas representaciones E-P en otra serie de funciones representadas en función de contar con 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".

En el ejemplo recogido en la siguiente Ilustración 3.8, en ella se observa una representación de 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].

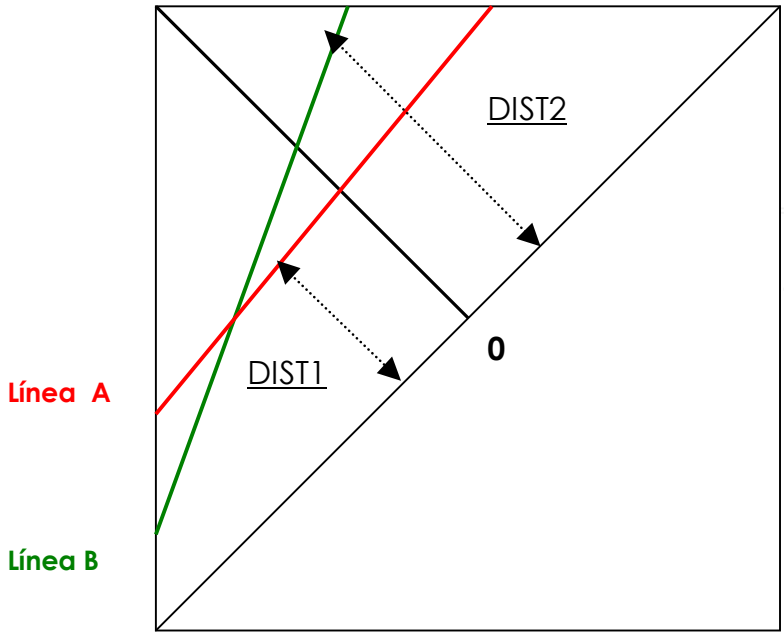


Ilustración 3.8 Ejemplo del Modelo de Swets. Fuente: Salton, G. and Mc Gill, M.J. Introduction to Modern Information Retrieval. New York: Mc Graw-Hill Computer Series, 1983..

El principal inconveniente de esta medida reside en el siguiente aspecto: "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. Asimismo, el valor de E[59] no puede obtenerse a partir de un simple umbral que distinga documentos relevantes o no relevantes de entre los documentos recuperados, 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].

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

Esta opinión de Salton responde plenamente a 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* como una aproximación logística a la estimación de los valores de *exhaustividad* y *precisión, "Robertson*, en colaboración con Teather, ha desarrollado un modelo para estimar las probabilidades correspondientes a E-P. Este procedimiento resulta inusual ya que al calcular una estimación de ambas probabilidades para una pregunta simple, toma en consideración dos cosas: la primera, la cantidad de datos empleados para alcanzar esas estimaciones y, segundo, 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 cuáles de aquellos son los relevantes" [RIJ, 1999]. Para Cooper este "ahorro de esfuerzo" es lo que se debe medir y para ello sólo haría falta una medida simple. Esta medida, en general, sólo se aplicaría a los sistemas que mostraran la salida de los documentos ordenados según un determinado 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] se recurre 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 operación booleana disyuntiva (operador "OR" generalmente de 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[60]. 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 la siguiente suma de 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 que necesita el usuario)

[60] Debe considerarse que en un extremo puede aparecer sólo un documento en cada subconjunto, mientras que en otro extremo pueden llegar a aparecer todos los documentos en el mismo subconjunto.

está en la primera posición, tres situaciones en segunda posición, 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]. 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 una valor simple obtenido a partir de 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[61]	12,49/7 = 1.78

Tabla 3.8 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 *exhaustividad* y *precisión* reside en la secuencialidad de la lectura de los resultados de una búsqueda: "los SRI típicos presentan los resultados al usuario formando una

[61] En este caso nos referimos a la longitud promedio esperada de búsqueda

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]. En la práctica, 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 muy común 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, 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, si la presencia de los documentos no relevantes al principio de la secuencia de documentos entregados al usuario es masiva, la sensación de frustración va a ser de cierta consideración, independientemente de que en el global de la búsqueda se le entreguen muchos más documentos relevantes que no relevantes. 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 [RIJ, 1999], [KOR, 1997], quien define una *exhaustividad normalizada* y una *precisión normalizada* para sistemas que presentan los documentos alineados según un determinado criterio de clasificación y donde no afecte el tamaño de la muestra analizada.

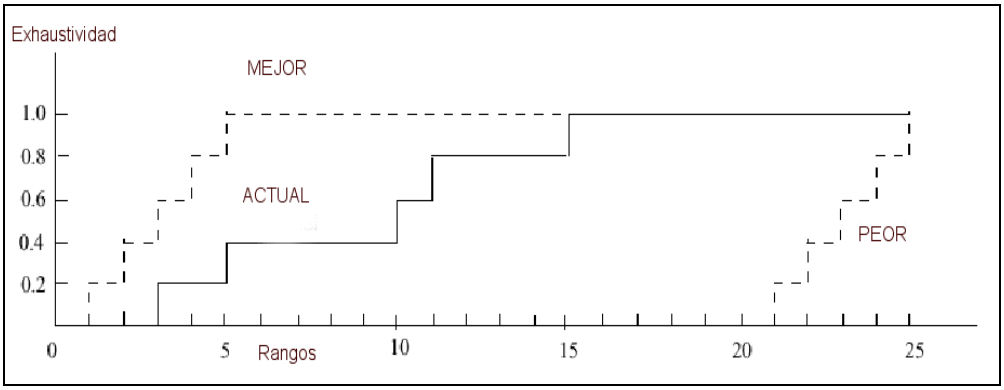


Ilustración 3.9 Ejemplo de cómo la *Exhaustividad* normalizada queda comprendida entre el pero y mejor resultado posible. Fuente: Rijsbergen, C.J. Information Retrieval. [En línea]. Glasgow, University, 1999. <http://www.dcs.gla.ac.uk/~iain/keith/> [Consulta: 21 de marzo de 2002]

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". 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 3.9.

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 se analice un documento relevante, el valor de la *exhaustividad* aumentará, hasta llegar a la unidad (en este caso, en el documento 15), permaneciendo constante hasta el último documento recuperado, ya que todos los documentos relevantes ya han sido recuperados. La comparación con la gráfica de la mejor búsqueda posible (correspondiente al caso de presentar los 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 refleja 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))[62]. 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,

[62] Siendo, en este caso, n1 el número de documentos relevantes y N el número total de documentos

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.

[SAL, 1983] y [KOR, 1997] presentan una nueva medida: la *ratio de deslizamiento*. Esta medida es bastante similar conceptualmente a la *exhaustividad normalizada* aunque presenta algunas diferencias: "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" [SAL, 1983].

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. 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 3.8 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 3.8 Ejemplo de determinación de la ratio de deslizamiento. Fuente: Korfhage, R.R. Information Retrieval and Storage. New York: Wiley Computer Publisher, 1997.

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" [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 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.

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 3.9 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[63]: " [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

[63] La representación de la tabla propuesta por Korfhage es la Tabla 3.9, es decir la inmediatamente anterior a este párrafo.

anterior en la siguiente, que Korfhage denomina: "tabla de diferencias de áreas" (Tabla 3.10).

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 3.10 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 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" [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 $\mathbf{I_1 = E + P}$ (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 I_1, 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 I_1:

Búsqueda	E	P	I_1
a	0.5	0.5	1
b	0.4	0.597	0.997
C	0.01	1	1.01

Tabla 3.11 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]. Esta deducción presenta un problema 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 I_2 [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]. Siguiendo con el ejemplo planteado, si se aplicara esta nueva medida a las tres búsquedas realizadas, los resultados serían:

Búsqueda	E	P	I_2
A	0.5	0.5	0.25
B	0.4	0.597	0.2388
C	0.01	1	0.01

Tabla 3.12 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 I_2 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 I_1 y I_2, 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].

Ejemplo práctico de cálculo de medidas de valor simple.

En la práctica, la medida I_1 de Borko y la medida I_2 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_1	I_2	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 3.13 Ejemplo del cálculo de las medidas de valor simple estudiadas para el ejemplo propuesto. Fuente: elaboración propia.

Estos datos confirman que normalmente todas las medidas responden similarmente, indicando que la mejor búsqueda es B3 (I_1 de Borko, I_2 de Voiskunskii y medida de Meadow, aumentan su valor frente a los valores de B1 y B2; mientras que las medidas de Heine y Vickery disminuyen su valor en B3 frente a las otras dos búsquedas). Esta coincidencia permite afirmar, que la efectividad de una 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.

Tablas e Ilustraciones.

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 67

Tabla 3.2 Resumen de las medidas empleadas en la evaluación convencional de la recuperación de la información. Fuente: Meadow, C. T. Text Information retrieval Systems. San Diego: Academic Press, 1993. 72

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..... 73

Tabla 3.3 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] 74

Tabla 3.4 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] 74

Tabla 3.5 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 75

Ilustración 3.2 Evolución típica de la *precisión* y de la *exhaustividad* en un SRI 76

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. 79

Tabla 3.6 Cálculo de pares de valores E-P de la búsqueda ejemplo. Fuente: elaboración propia.82

Ilustración 3.4 Evolución distribución pares de valores E-P del ejemplo. Fuente: elaboración propia. 83

Ilustración 3.5 Evolución distribución valores E-P del ejemplo corregida. Fuente: elaboración propia. 84

Tabla 3.7 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. 84

Ilustración 3.6 Representación de la evolución E-P real y óptima del ejemplo. Fuente: elaboración propia. 85

Tabla 3.8 Medidas de la calidad de una búsqueda propuestas por varios autores. Fuente: Meadow, C. T. *Text Information retrieval Systems*. San Diego: Academic Press, 1993..... 87

Ilustración 3.7 Representación de la evolución de las medidas de calidad de una búsqueda ejemplo. Fuente: elaboración propia. 88

Ilustración 3.8 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.. 90

Tabla 3.8 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, 1997.93

Ilustración 3.9 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] 94

Tabla 3.8 Ejemplo de determinación de la ratio de deslizamiento. Fuente: Korfhage, R.R. Information Retrieval and Storage. New York: Wiley Computer Publisher, 1997. 97

Tabla 3.9 Ejemplo propuesto por Korfhage para determinar la medida de la satisfacción. Fuente: Korfhage, R.R. Information Retrieval and Storage. New York: Wiley Computer Publisher, 1997.98

Tabla 3.10 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. 99

Tabla 3.11 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. 101

Tabla 3.12 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. 101

Tabla 3.13 Ejemplo del cálculo de las medidas de valor simple estudiadas para el ejemplo propuesto. Fuente: elaboración propia. 103

4 La evaluación de los SRI en la web.

RESUMEN: La necesidad crítica de evaluación de los SRI subyace tras su traslado al contexto de la web. En este nuevo contexto han surgido nuevos tipos de sistemas y nuevos problemas que han dado lugar, al desarrollo de una serie de trabajos específicamente concebidos para analizar la viabilidad de la recuperación de información. En estos trabajos destacan tres grandes grupos: los dedicados a analizar las características formales o externas del sistemas, los que someten a estos sistemas a algún tipo de ensayo o experimentación y finalmente, aquellos trabajos orientados a proponer una metodología científica para su evaluación.

La evaluación de los SRI en la web.

Introducción.....107

Estudios explícitos.....109

Aspectos formales y prestaciones en la recuperación de información. 110
Tamaño del índice del motor de búsqueda.....111
Audiencia del motor. 117
Número de consultas realizadas.....119
Tráfico redirigido.....120
Porción de la página indexada.....121
Fidelidad. 122

Sinopsis de estudios explícitos.....122

Estudios experimentales.....124

Revisión de estudios realizados.....124

Chu y Rosenthal (1996).....124
Leighton y Srivastava (1995-1999). 126
Arents (1995-2001). 127
Bharat y Broder (1997). 128
Ralph (1997). 128
Wishard (1998). 129
Lawrence y Giles (1998). 129
Gordon y Pathak (1999). 130
Gwizdka, J. and Chignell (1999). 134
Ming (2000). 136
Notess (2000). 137
Ljoslund (1999-2000). 138
CNET search site olympics (Thomas, 2002). 143

Sinopsis de estudios experimentales.....143

Propuestas globales de evaluación.....146

Oppenheim (2000).....147
Savoy y Picard (2001).....151
Schlichting y Nilsen (1997). 154
Johnson, Griffiths y Hartley (2001). 156

Tablas e ilustraciones. 163

Introducción.

El paralelismo existente entre el crecimiento de la información depositada en diversas fuentes dentro de la web y la proliferación del número de los motores de búsqueda que aseguran poseer la mayor cantidad de recursos debidamente indexados y accesibles a través de ellos, ha sido ya comentado con anterioridad.

Al poco tiempo de su aparición y de la popularización de su uso en el seno de los usuarios de la red de redes, se ha generado la siguiente incógnita: "¿cuál es el mejor de estos ingenios de búsqueda?". Casi todos los usuarios de Internet creen poseer la respuesta a esta pregunta. No obstante, si reflexionaran detenidamente antes de contestar, ¿cuáles son las bases sobre las que fundamentarían su respuesta?. El conjunto de posibles argumentos puede ser variado: familiaridad con un motor determinado, pleno convencimiento de su calidad, desconocimiento del resto, costumbre, etc. Sea cual fuere la respuesta, lo que está claro es que sobre estos nuevos SRI, sigue latente la necesidad de evaluación comentada al principio del capítulo precedente.

Con el objeto de otorgar una mayor certidumbre a esta respuesta, ha proliferado la producción científica en este campo. A este respecto, Oppenheim manifiesta: "los artículos que informan de la evaluación de los motores de búsqueda han ganado popularidad a medida que la web ha crecido" [OPP, 2000].

Los resultados que arrojan estos estudios ofrecen resultados *dispares* y *dispersos*. La *disparidad* proviene de la obtención de resultados distintos entre la mayor parte de los estudios, aunque siempre se pueden identificar algunas conclusiones comunes en todos ellos.

La *dispersión* surge cuando el conjunto de motores evaluados es funcionalmente distinto, lo que dificulta la especificación de conclusiones válidas y generales, al evaluar cada estudio un conjunto de motores que poco tiene que ver con el evaluado por otro. Adicionalmente, incide en esta dispersión el hecho de que estos estudios no suelen repetirse con el tiempo, lo que conlleva a que los mismos ofrezcan conclusiones válidas sobre una situación determinada, pero incapaces de percibir la evolución de esta tecnología, algo verdaderamente preocupante considerando los continuos cambios que se producen en este campo, tanto de carácter técnico como de incremento de información.

Profundizando en la disparidad de los estudios realizados, también se presentan grandes diferencias relacionadas con el método empleado en su

realización o con su alcance. Existe una considerable cantidad de trabajos donde el proceso de evaluación se limita a los aspectos explícitos (también se podrían denominar "externos", "formales" o "testimoniales") del motor de búsqueda: amigabilidad de la interface, velocidad de respuesta, formatos de presentación, documentación existente, ayuda del sistema, etc. Evidentemente, todos estos aspectos poseen su importancia, pero no interpretan, en modo alguno, la efectividad de las operaciones de recuperación implementadas[64].

Otro amplio conjunto de trabajos sí está centrado en el estudio de esta efectividad haciendo uso, en la mayoría los casos, de las medidas *exhaustividad* y *precisión*.

Es decir, al igual que en los SRI tradicionales, se hace uso de los juicios basados en la *relevancia* [LAN, 1973]. En el capítulo anterior, se ha reflejado suficientemente que estas medidas, por sí solas, tampoco pueden ofrecer una serie de conclusiones definitivas y concluyentes sobre las prestaciones de los motores de búsqueda y que, además, cuentan con el problema añadido de poseer ciertas dosis de subjetividad subyacentes. A pesar de todo, estos estudios sí arrojan algunos resultados mucho más convincentes que las evaluaciones meramente formales.

Al abordar el diseño de un estudio de este tipo, surgen problemas muy vinculados con la propia naturaleza de la web, presentándose situaciones que, en el entorno de los SRI tradicionales, resultarían de difícil imaginación. Esto se debe, entre otras muchas razones, a que el propio origen de la web poco tiene que ver con la estructura de un SRI, sino más bien con la de un sistema orientado a la difusión de información y de trabajo en grupo. Oppenheim presenta varios de estos problemas: presencia de registros duplicados, tendencia a recuperar documentos poco relevantes, recuperación de enlaces que llevan a páginas web que ya no existen (o que han sido trasladadas a otra ubicación), el "spamming" (o uso de técnicas inadecuadas por parte de los autores de las páginas para atraer visitas), criterios de alineación de los documentos recuperados inadecuados, etc. No obstante, para este autor, "el problema fundamental de la web es su tamaño, heterogeneidad e inconsistencia; los recursos cambian y se multiplican y la impresionante velocidad de respuesta de los motores de búsqueda no está ligada, generalmente, a una efectiva recuperación de información" [OPP, 2000].

---

[64] Aunque no es lo más usual, algunos estudios se limitan a una descripción de las características de los motores de búsqueda sin ninguna aportación valorativa.

La naturaleza de los SRI en la web, ciertamente atípica, propicia que no puedan emplearse únicamente las medidas basadas en la *relevancia* de los documentos recuperados, sino que además es necesario el establecimiento de un grupo de medidas específicas, tal como es el caso de la ratio de *enlaces fallidos*, es decir, el porcentaje de los documentos que un motor de búsqueda entrega como respuesta y que se encuentran inactivos (las páginas a las que dirigen ya no existen, o se encuentran en otra ubicación o la referencia contiene errores ortográficos que remiten a una página inexistente). El *solapamiento* y el *acierto único* son dos medidas que cobran aquí especial significado, aunque también han sido utilizadas en evaluaciones de SRI clásicos [MEA, 1993], son mucho más profusamente empleadas en este entorno.

El cálculo de este conjunto de medidas unido, generalmente, a la determinación de la *cobertura*, la *precisión* y la *exhaustividad*, conforma un segundo grupo de estudios, que se pueden denominar como *estudios experimentales*, en tanto que llevan a cabo algún ensayo antes de mostrar sus conclusiones.

Finalmente, es frecuente encontrar algunos sitios web que recogen recomendaciones para el uso de motores de búsquedas basadas en las características generales de cada motor o en la experiencia de los autores de las mismas. En este grupo de estudios aparecen recomendaciones del siguiente tipo: "si deseas localizar documentos con palabras raras, usa *Alta Vista*", "si deseas hacer preguntas en lenguaje natural usa *Northern Light*", etc.

Estos estudios resultan de difícil encuadre en cualquiera de los dos grupos anteriores, debido a que su carácter de recomendación no se corresponde ciertamente con una simple descripción de las prestaciones de cada motor o con el resultado de un cálculo de una medida determinada. No obstante, deben tomarse en consideración, en tanto que son unas interesantes guías de referencia del uso de estos SRI, especialmente indicadas para los usuarios no iniciados por su carácter divulgativo. Es por ello que resulta apropiado englobarlos dentro de un tercer grupo, al cual podría denominarse como *divulgativos*.

Estudios explícitos.

Los primeros estudios de evaluación de los motores de búsqueda datan de 1995, aproximadamente un año después de la aparición de los primeros

motores. En un estudio que, con el paso del tiempo, se está convirtiendo en un clásico en este campo, Chu y Rosenthal [CHU, 1996], de la Universidad de Long Island en Nueva York, recopilan parte de ellos y presentan sus conclusiones en la Conferencia Anual de ASIS de 1996.

Aspectos formales y prestaciones en la recuperación de información.

Los primeros estudios citados por estos autores se encuentran basados, en la mayoría de los casos, en las características formales del propio motor de búsqueda y en las descripciones técnicas que proporciona el sistema[65]. En su trabajo mencionan a Courtois, Baer y Stark, quienes destacan la potencialidad del motor *Webcrawler* en todo lo relacionado con su flexibilidad a la hora de plantear las ecuaciones de búsqueda y su rápida respuesta; resaltando, además, su interface, la cual consideran muy adecuada para usuarios poco iniciados.

Basándose también en la idea de la flexibilidad de la interface, Scoville apuesta por los motores *Excite*, *Infoseek* y *Lycos*. Chu y Rosenthal citan también el trabajo de Kimmel, quien califica a *Lycos* como el mejor a partir del estudio de la documentación aportada por varios motores y de sus características externas. De similar naturaleza es el estudio de Davis, quien considera el tamaño del índice del motor y las posibilidades de recuperación de información, decantándose por *Alta Vista*, *Hot Bot* e *Infoseek*, sobre un total de siete motores evaluados [DAV, 1996].

Westera analiza las capacidades de la interface de los motores de búsqueda a la hora de realizar las operaciones de recuperación de información. Una particularidad interesante de este estudio, es que el mismo se actualiza periódicamente desde 1996 (resalta la inserción de *WISEnut* en la última revisión del estudio). Divide las capacidades de búsqueda en dos grupos: básicas y especiales. En el primer grupo destacan *Alta Vista* y *Google*; en el segundo *Alta Vista* y *Hot Bot* [WES, 2001].

Chu y Rosenthal comentan que *C|net,* empresa especializada en evaluar productos de servicios de información en línea, realizó un estudio comparativo de diecinueve motores de búsqueda tomando en consideración el acierto en sus búsquedas, la facilidad de uso del sistema y la cantidad de opciones avanzadas que proporcionaba cada ingenio. En este estudio se destaca a *Alta Vista* como el mejor de todos [CHU, 1996].

---

[65] Esta enumeración de estudios realizados les sirve de preámbulo para presentar el suyo propio, el cual se encuadra en el grupo de estudios experimentales.

En 1996, Slot analiza exhaustivamente los contenidos manejados por dieciséis motores de búsqueda y directorios, considerando también objeto de su estudio, entre otras variables, el tiempo de respuesta y la claridad de la interface. En un segundo nivel de su estudio analiza con profundidad las posibilidades que cada motor ofrece en la recuperación de información. En el primer apartado de su estudio, considera a *Alta Vista* y *Yahoo* como excelentes, seguidos muy de cerca por *Lycos*. En el segundo apartado sigue destacando *Alta Vista* ligeramente sobre los demás [SLO, 1996].

También en el año 1996, Zorn, Emanoil y Marshall comparan cuatro motores de búsqueda (*Alta Vista*, Infoseek, *Lycos* y Opentext), analizando su número de documentos estimado, su documentación, qué porción de página indexan, las posibilidades de búsqueda de cada uno, si detectan documentos duplicados en el índice y, por último, si emplean un método de alineamiento en la presentación de los documentos [ZOR, 1996]. Los resultados de este estudio presentan a *Alta Vista* y *Lycos* como los motores de mayor tamaño (índices casi diez veces más grandes que los de los otros), gran igualdad en prestaciones de búsqueda y formato de presentación. La documentación aportada por *Alta Vista* y Opentext se considera excelente y todos indexan la página completa a excepción de *Lycos*.

Tamaño del índice del motor de búsqueda.

En un contexto ideal, si un motor de búsqueda recopilara en su índice la totalidad de los documentos de la web (o un porcentaje cercano), sin duda alguna, ese motor sería el predilecto de todos los usuarios de Internet, otorgando a este parámetro un valor prioritario por encima de otros, quizás mucho más importantes, tales como son la *relevancia* de los documentos con respecto a la pregunta, la presencia de documentos duplicados o la inclusión de enlaces erróneos en el índice. La realidad es bien distinta, "el constante cambio y expansión de la web, provoca que ninguno de los motores de búsqueda pueda indexar la totalidad de sus documentos. Muchos estudios, diseñados para estimar el tamaño de la web, han determinado que los motores poseen recopilados entre el 5% y el 30% de la totalidad de documentos de la web, y la unión de los once principales motores de búsqueda no alcanza el 50%" [CHA, 2001]. Delgado Domínguez afirma que "la intersección de los motores de búsqueda más conocidos es tan sólo de un dos por ciento" [DEL, 1998], así sólo un subconjunto de páginas muy específico de la globalidad de la web forma parte de los índices de los principales motores y que las diferencias entre ellos son muy grandes, debido, seguramente a sus distintos algoritmos de rastreo.

No obstante, la estimación del tamaño de los índices y la determinación (aproximada) del porcentaje de la web recogido en sus índices, ha sido objeto de interés de varios autores. Uno de los trabajos más citados es el realizado por Lebedev en 1996, quien desarrolló un sencillo experimento consistente en realizar ocho preguntas a distintos motores. Cada pregunta la forma sólo una palabra clave, relacionada con la Química o con la Física. Posteriormente procedió a la suma del total de los documentos devueltos por cada motor en cada una de las ocho búsquedas. Lebedev considera que el número de documentos es el indicador más primario que puede utilizarse (pero indicador al fin y al cabo), para ponderar la calidad de los distintos motores de búsqueda. Lebedev también sugiere como segundo criterio la incorporación de nuevas páginas al motor, factor al que denomina "dinámica de los motores de búsqueda", concluyendo desde su punto de vista que *Alta Vista* es el mejor motor de búsqueda, seguido muy de cerca por *HotBot*. Llama la atención, en palabras del autor de este experimento, el pésimo resultado de *Yahoo*, *Galaxy* y *Webcrawler* [LEB, 1997].

Slot también analizó el tamaño del motor o directorio, resaltando el inmenso tamaño de *Alta Vista* frente a los demás analizados. Como muestra destaca que calculó en veintiún millones de documentos el tamaño de este motor, seguido de *Lycos* con cinco millones de documentos[66] [SLO, 1996]. El tamaño del motor, el número de documentos devueltos y una amplia serie de características relacionadas con el modo de almacenar las referencias y el formato de salida de los documentos, constituye el objeto del trabajo desarrollado por Peterson, quien analizó ocho motores de búsqueda: *Alta Vista*, *Excite*, *Hot Bot*, *Infoseek*, *Lycos*, *Opentext*, *Ultra* y *Webcrawler*. Para estimar el tamaño de los motores y precisar cuál de ellos devuelve más documentos, se tomaron datos en tres períodos de tiempo distintos a lo largo del año 1996 (febrero, mayo y noviembre), interrogando a estos ocho sistemas por medio de dos expresiones de búsqueda, una conformada por un término individual y la otra por una frase literal. El resultado de esta parte del experimento mostraba al motor *Hot Bot* como el de mayor número de documentos gestionados, superando ligeramente a *Excite* y *Ultra*. El análisis de las características de la presentación de los documentos y de cómo se almacenan los documentos, muestra también a *Hot Bot* como el motor más completo, junto con *Alta Vista* [PET, 1997].

---

[66] Estos valores están completamente desfasados en la actualidad, aunque, en el contexto en el que fueron calculados poseían la suficiente vigencia para recordarlos.

El número de documentos contenido en cada motor y sus posibilidades de búsqueda, son las características analizadas por Maldonado y Fernández en su estudio de 1998. En el mismo se empleó una utilidad presente en los motores de búsqueda *Alta Vista* e *Infoseek*, la cual permite recuperar páginas que apuntan (enlazan) a una determinada dirección. El cómputo final de los resultados situaba a un directorio, *Yahoo*, como el sistema que más páginas tenía recogidas en sus índices, seguido (por este orden) de los motores: *Alta Vista*, *Excite*, *Lycos*, *Webcrawler*, *Hot Bot*, *Magellan*, *Infoseek*, *Galaxy* y *Looksmart*. En el análisis de las posibilidades de búsqueda, se encontraban grandes similitudes entre *Alta Vista* e *Infoseek*, aunque se decantan ligeramente por el primero [MAL, 1998].

Es precisamente el tamaño de los índices que gestionan los distintos motores, una de las variables a analizar más susceptibles de obsolescencia (sino la que más), ante la naturaleza dinámica de estos sistemas (que continuamente incorporan nuevos documentos a sus índices) y la introducción de nuevos motores que, de forma vertiginosa aumentan su número de usuarios y el volumen de documentos gestionados.

Para disponer de una adecuada visión de la situación en torno a este parámetro, es necesario consultar estudios actualizados con cierta frecuencia. Bajo esta perspectiva, destacan los trabajos de Notess, quien realiza evaluaciones del tamaño de los índices desde casi el nacimiento de estos SRI. Entre octubre de 1996 y mayo de 1998 (fechas contemporáneas a las de los estudios anteriormente citados), *Alta Vista, Northern Light* y *Hot Bot* comparten los puestos de cabeza en esta clasificación, tal como podemos ver en la siguiente tabla [NOT, 1998].

Fecha	**1**	**2**	**3**
May 1998	*Alta Vista*	HotBot	*Northern Light*
February 1998	HotBot	*Alta Vista*	*Northern Light*
October 1997	*Alta Vista*	HotBot	*Northern Light*
September 1997	*Northern Light*	*Excite*	HotBot
June 1997	HotBot	*Alta Vista*	*Infoseek*
October 1996	HotBot	*Excite*	*Alta Vista*

Tabla 4.1 Clasificación de los motores de búsqueda según su tamaño. Fuente: Notess, G. R. Comparing Internet search engines [En línea]. Charles Sturt University, 1999. <http://www.csu.edu.au/special/online99/proceedings99/103a.htm> [Consulta: 10 noviembre 2001]

Notess prosigue, en la actualidad su línea de trabajo en la web *Searchengine Showdown*[67] [NOT, 2000], donde publica, entre otros, el número de documentos recuperados por distintos motores tras la realización de veinticinco consultas. En agosto de 2001 proporcionaba la siguiente clasificación: *Google*, *All the Web*, *WISEnut*, y *Northern Light*.

Resulta un hecho revelador de la evolución de la web, que en poco más de tres años, tres motores se habían aupado a las primeras posiciones de esta clasificación, superando el tamaño de su índice a los primeros de la tabla anterior.

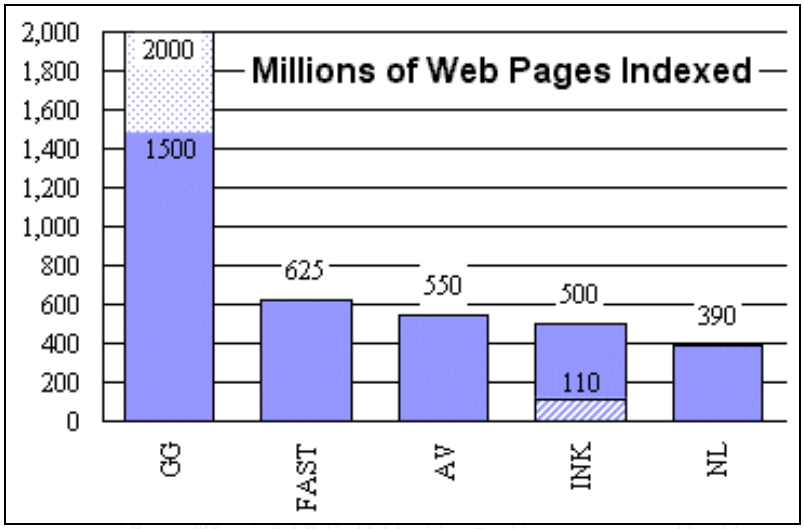
Los mismos motores permanecen a la cabeza en la actualización de estudio realizada a principios de marzo de 2002, tal como refleja la Tabla 4.2. También se muestra la discrepancia existente, a la hora de estimar el tamaño del índice de cada motor, entre Notess y la publicidad que llevan a cabo los motores (casi todos suelen proporcionar un volumen aproximado de documentos en sus índices).

Motor	**Estimación**	**Longitud anunciada**	**% de diferencia**
Google	968	1.500	54,9
WISEnut	579	1.500	159
Altheweb (Fast)	580	507	-13,5
Northern Light	417	358	-15,2
Altavista	397	500	25,9
Hotbot	332	500	50,6
MSN Search	292	500	71.23

Tabla 4.2 Clasificación actualizada de los motores de búsqueda según tamaño y determinación de diferencias entre tamaños estimados y tamaños declarados. Fuente: Search Engine Statistics: Database Total Size Estimates. [En línea]. Bozeman, MT: Notes.com, 2000. <http://www.searchengineshowdown.com/stats/sizeest.shtml/> [Consulta: 12 marzo 2002]

[67] Proporciona estadísticas adicionales usando medidas que han encontrado en la web un campo extenso de aplicación: el solapamiento, el acierto único y el porcentaje de enlaces fallidos.

En la web *Searchenginewatch*, Sullivan presenta también su estimación del tamaño del índice de los motores de búsqueda, que representa en la gráfica que recogemos en la Ilustración 4.1.



GG=*Google*, **FAST**=FAST, **AV**=AltaVista **INK**=Inktomi **NL**=Northern Light.

Ilustración 4.1 Clasificación de los motores de búsqueda según tamaño. Fuente: Search Engine Sizes. [En línea]. Darien, CT: INT Media Group, 2001 <http://www.searchenginewatch.com/reports/sizes.html> [Consulta: 10 de febrero de 2002]

El estudio, realizado en diciembre de 2001, sitúa a *Google* en primer lugar destacado frente a los demás motores y a *Fast* en segundo lugar.

Es decir, este estudio exhibe resultados coincidentes con el anterior, con la excepción del motor *WISEnut*. El autor indica que llama la atención el hecho de que *Google* aparezca con dos cifras distintas, pero que esto se debe realmente a que "*Google* sólo ha indexado 1 billón y medio[68] de páginas, pero a causa del uso extensivo de los enlaces que realiza este motor, puede actualmente devolvernos listas de páginas adicionales que nunca ha visitado" [SUL, 2001a].

Otro caso particular es el del motor *Inktomi*, sistema que diferencia entre dos grandes conjuntos de sitios web recopilados; el primero de ellos es el denominado "Lo mejor del web", que agrupa a 110 millones de documentos

[68] A la hora de valorar esta cifra, hay que recordar que los norteamericanos consideran que 1000 millones es 1 billón.

y el segundo grupo, denominado "El resto del web", que agrupa 390 millones de páginas[69].

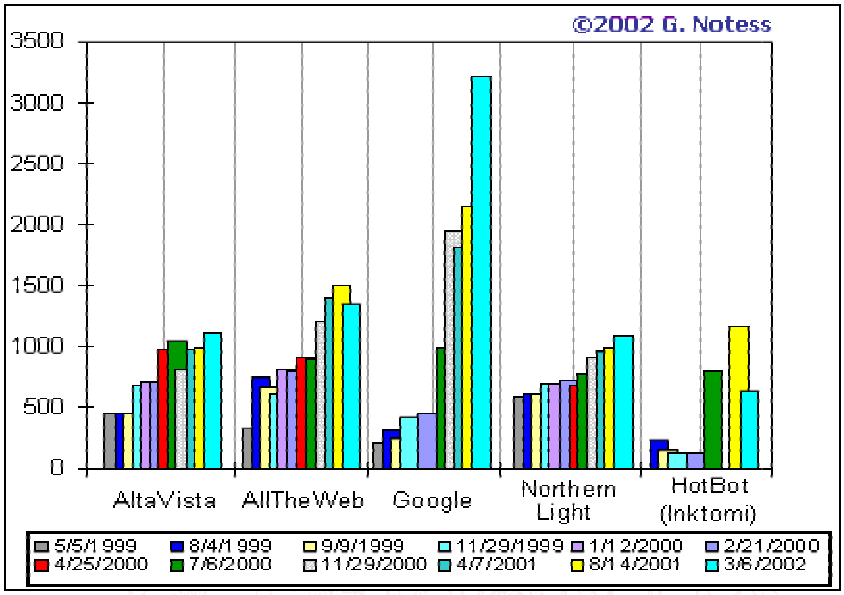


Ilustración 4.2 Evolución en el tiempo del tamaño de los motores de búsqueda. Fuente: Search Engine Database Change Over Time. [En línea]. Bozeman, MT: Notes.com, 2000 <http://www.searchengineshowdown.com/stats/change.shtml> [Consulta: 10 de marzo de 2002]

En este informe también se analiza la evolución de esta estimación a lo largo del tiempo, en los informes realizados desde el año 1998 era *Alta Vista* el motor de mayor tamaño, siendo igualado en junio de 2000 por *Fast* y posteriormente rebasados ambos en septiembre de 2000 por la vertiginosa ascensión de *Google*. Esta evolución es similar a la presentada por Notess [NOT, 2000], quien también analiza el porcentaje de crecimiento de los motores en relación con su último estudio. Los datos de marzo de 2002 reflejan que *Google* es el motor de mayor crecimiento desde agosto de 2001, seguido de WISEnut, *Alta Vista* y Northern Light; por el contrario AllTheWeb, HotBot y MSN (entre otros) han tenido un crecimiento más pequeño.

---

[69] Para ampliar información sobre el motor Inktomi se recomienda la consulta del documento: '*Inktomi Reenters Battle For Biggest*', sito en la página web <http://www.searchenginewatch.com/sereport/00/06-inktomi.html>.

Audiencia del motor.

Otro criterio explícito, más empleado de lo que podría suponerse, es el nivel de audiencia del motor de búsqueda. En la web *Searchengineswatch.com* figuran varios trabajos de Sullivan, quien sintetiza los datos procedentes de organizaciones dedicadas a medir esta audiencia. El primero de ellos es el elaborado por *Jupiter Media Matrix*[70], a partir de los accesos que más de cien mil usuarios de Internet realizan sobre los motores de búsqueda o directorios analizados. Este informe proporciona como motor más popular a *Yahoo* (64.8%) seguido por *Microsoft Network* (58.3%) y *American Online* (46%) [SUL, 2001b].

Si bien este estudio viene avalado por su gran y continua interactividad con los usuarios, y por su actualización, introduce una considerable distorsión en sus resultados, en tanto que muchos de los sistemas analizados no son verdaderamente motores de búsqueda, sino que se tratan más bien de directorios. No obstante, se pueden extraer varias conclusiones importantes a partir de los datos ofrecidos por este estudio. La primera es que los tres sistemas mencionados muestran una perceptible tendencia hacia el crecimiento de su audiencia si los comparamos con los datos del mismo estudio realizado hace un año aproximadamente[71]. Otro dato interesante que aporta este estudio, es la clara evolución creciente de *Google* y el descenso vertiginoso que han sufrido los accesos a *Alta Vista* (Ilustración 4.3). Como factor que propicia el ascenso de *Google*, además de la gran popularidad que está alcanzando entre sus usuarios y las dosis de confianza que estos van adquiriendo en este motor, Sullivan destaca también los aspectos positivos de su colaboración con un directorio: *Netscape*.

---

70 *Jupiter Media Matrix* es una empresa especializada en medir la audiencia a los sitios web y el impacto de Internet y las nuevas tecnologías en el comercio electrónico.

71 Martínez Méndez, F.J. "Aproximación general a la evaluación de la recuperación de información por medio de los motores de búsqueda en Internet". *Scire*, vol 7, nº 1, 2001. También accesible [en línea] <http://www.um.es/gtiweb/fjmm/ibersid2000.PDF> Universidad de Murcia, 2001. [Consulta: 19 de octubre de 2001]

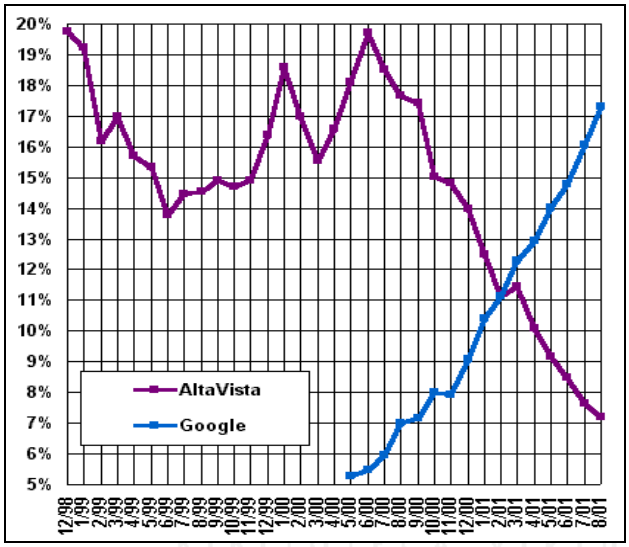


Ilustración 4.3 Evolución en el tiempo del tamaño de los motores *Alta Vista* y *Google*. Sullivan, D. Media Metrix Search Engine Ratings [En línea]. Darien, CT: INT Media Group, 2000 [Consulta: 10 de febrero de 2002]

Sullivan analiza otro estudio de audiencia interesante, el realizado por *Nielsen/NetRatings*[72]. Este estudio, de actualización semanal, presenta la novedad de medir el porcentaje de usuarios que han visitado los distintos sitios, el tiempo medio que una persona ha dedicado a este motor de búsqueda, además de medir la audiencia según el número de accesos computados.

La Tabla 4.4 recoge el nivel de audiencia de los doce sitios más visitados en el mes de octubre de 2001, basado en los datos proporcionados por más de 62.000 usuarios que se conectan desde sus casas [NET, 2001]. No todos los sitios web que presenta este análisis son motores de búsqueda, sino que también incorpora directorios y sitios web de otra naturaleza, tales como las web de Microsoft, Amazon o Disney, por ejemplo. La serie de datos ofrecidos sitúa a *Yahoo* en primer lugar (48%) y a *Microsoft Network* en segundo lugar (38,5%). El primer motor de búsqueda que encontramos es *Google* (16,2%), situado en tercer lugar, seguido de *Lycos* en cuarto lugar con algo más de un once por ciento [SUL, 2001c].

[72] *Nielsen Media Research* es la más importante de las compañías norteamericanas especializadas en el cálculo de índices de audiencia en los medios de comunicación. *NetRatings, Inc*. es un proveedor de información en Internet que opera en 26 países.

	Audiencia	Alcance	Tiempo medio
1. AOL Time Warner	66,000,216	63.64	0: 56: 40
2. *Yahoo*!	56,905,756	54.87	1: 33: 44
3. MSN	51,636,141	49.79	1: 19: 34
4. Microsoft	29,027,086	27.99	0: 09: 52
5. *Lycos* Network	24,033,326	23.17	0: 15: 22
6. Amazon	17,733,580	17.10	0: 14: 58
7. Walt Disney Internet Group	17,461,763	16.84	0: 30: 23
8. About-Primedia	17,205,406	16.59	0: 10: 03
9. eUniverse Network	16,828,229	16.23	0: 16: 56
10. eBay	16,341,325	15.76	1: 40: 49
11. *Google*	14,557,460	14.04	0: 20: 21
12. *Excite*@Home	14,536,676	14.02	0: 47: 54

Tabla 4.4 Doce de los veinticinco sitios más visitados de Internet. Fuente: NetRatings/Nielsen.

Este estudio confirma gran parte de los datos ofrecidos por el anterior, certificando el descenso de popularidad de *Alta Vista* y el ascenso progresivo de *Google*, aunque destacan los buenos resultados que obtiene *Lycos* en estudio, comparados con el anterior.

Número de consultas realizadas.

En otro trabajo, Sullivan vuelve a proporcionar una clasificación del número de consultas que reciben cada día los motores de búsqueda. La Tabla 4.5 es producto del análisis de diversas fuentes que miden el uso de los motores.

Esta nueva clasificación aporta un elemento adicional al número de visitas que recibe cada motor: la *utilidad* del mismo hacia sus usuarios. En este caso, aunque *Google* mantiene su posición de predominio, sorprende la segunda posición del motor Inktomi, hecho que puede llamar la atención si no se considera la incidencia que en esta medida tienen las alianzas estratégicas que se vienen desarrollando entre directorios, motores y portales de Internet.

Motor	Consultas
Google	130
Inktomi	80
Alta Vista	50
Direct Hit	20
Fast	12
Goto	5
Ask Jeeves	4

Tabla 4.5 Millones de búsquedas por día en distintos motores. Fuente: Searches Per Day. [En línea]. Darien, CT: INT Media Group, 2001 <http://www.searchenginewatch.com/reports/perday.html> [Consulta: 12 de marzo de 2002]

En este caso particular, el total de consultas de *Google* responde a las efectuadas en su propio sitio web y las que se computan a partir de sus sitios afiliados (*Nestcape* y *Yahoo*), que redirigen una importante cifra. Situación similar ocurre con *Inktomi*, cuyo motor es empleado por una gran cantidad de portales genéricos de Internet o *Direct Hit* (que agrupa las visitas de varios portales afiliados). La estadística de accesos de *Fast* refleja, en cambio, sólo las consultas realizadas en su sitio web, por lo que ocupa un discreto lugar en esta clasificación y computa muchas menos consultas que los anteriores [SUL, 2001d].

Tráfico redirigido.

En un tercer trabajo, Sullivan muestra una clasificación de los motores de búsqueda y directorios, a partir del tráfico que estos sistemas han generado hacia los distintos sitios web a los que remiten. Este estudio no mide la popularidad de la fuente de información, sino la utilidad de la misma como referencia al sitio web donde se aloja el documento de interés.

En este caso, Sullivan toma sus datos del informe *StatMarket Search Engine Ratings,* estudio que analiza los registros de accesos de múltiples sitios web (más de cien mil) para verificar de dónde proceden las visitas recibidas. El resultado de este estudio es muy explícito a favor de *Yahoo*, tal como muestra la Ilustración 4.4:

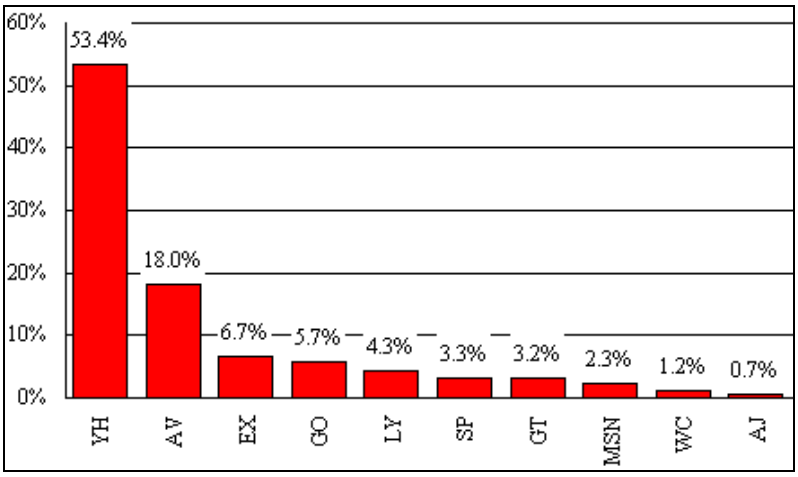


Ilustración 4.4. Accesos de múltiples sitios web. Fuente: StatMarket Search Engine Ratings. [En línea]. Darien, CT: INT Media Group, 2000. < http://www.searchenginewatch.com/reports/statmarket.html/> [Consulta: 24 de noviembre 2001]

Este estudio repite la evaluación conjunta de motores de búsqueda y directorios, y de nuevo, *Yahoo* aparece muy destacado en cuanto a nivel de audiencia y de utilidad, aunque, a diferencia de los anteriores trabajos de Sullivan comentados, predominan los accesos a motores de búsqueda sobre los accesos a directorios ya que, tras *Yahoo*, aparecen *Alta Vista*, *Excite*, *Google*[73] y *Lycos* [SUL, 2000].

Porción de la página indexada.

En un estudio de 1995, Winsip estudia cuatro motores de búsqueda: *WWWWorm*, *Webcrawler*, *Lycos* y Harvest; junto a dos directorios: *Yahoo* y *Galaxy*, centrando su análisis en la porción de página indexada por cada uno de ellos, en el interface del sistema, en las capacidades de búsqueda, en los formatos de presentación de los documentos y en el número de documentos recuperados, destacando en primer lugar a *Lycos*, ligeramente por encima de *Harvest* [WIN, 1995].

[73] Llama la atención la discreta posición de *Google*, debida a la fecha de realización del estudio, que casi coincide con la puesta en marcha de este motor de búsqueda.

Fidelidad.

En 1996, en la *Universidad de Sunderland*, Stobart y Kerridge desarrollaron un estudio sobre un conjunto de cuatrocientos usuarios de Internet (profesores e investigadores universitarios en su mayor parte), que cumplimentaron un cuestionario donde indicaban cuáles eran los motores de búsqueda que más empleaban (obteniéndose como respuesta *Alta Vista*, *Yahoo* y *Lycos*), de forma preferente; y en segundo lugar, se preguntaba a quienes aseguraban usar más de un motor de búsqueda cuál era al que acudían primero, siendo *Alta Vista* el preferido con una amplísima diferencia. En tercer lugar se interrogaba a los participantes en el estudio sobre las posibles causas de su fidelidad reconocida a un motor, y las tres principales razones (todas ellas en torno a un 20%) que formaron parte de la respuesta eran, por este orden: velocidad, tamaño del índice y costumbre [STO, 1996].

Sinopsis de estudios explícitos.

La revisión de los criterios empleados en la evaluación explícita de los motores de búsqueda, compendia veinte trabajos que han empleado trece variables diferentes de análisis. El número medio de características analizadas por estudio es bastante bajo, en tanto que sólo se han analizado 38 características en los 20 estudios (1.9 por trabajo). Las características más empleadas son: tamaño del índice (7 veces)[74], posibilidades de búsqueda (6 veces) e interface de usuario (5 veces). Las referencias a estos trabajos y la serie de variables empleadas en cada uno de ellos están recogidas en la Tabla 4.6.

La última columna de la tabla indica el motor que recibe mejor valoración en cada estudio. Hasta siete motores aparecen citados como los mejores según el objeto del estudio. Una agrupación con base en esta serie de características más estudiadas, produce coincidencias en sus conclusiones.

Así, en los estudios que analizan el tamaño del índice, los resultados eran favorables al motor *Alta Vista*, hasta que en los mismos se comenzó a reflejar la presencia de *Google*, siendo este motor el que ocupa la posición de privilegio desde entonces. Cuando los estudios se basan en el análisis de las posibilidades de búsqueda, es *Alta Vista* el más veces destacado y cuando

[74] Este número podría elevarse a ocho si consideráramos que en el segundo estudio citado de Notess se analiza el crecimiento del tamaño del índice.

los estudios analizan la interface gráfica de usuario, no existe una tendencia común en los resultados.

	PB	**TR**	**GUI**	**DO**	**TI**	**CRE**	**ACI**	**FA**	**AUD**	**TRA**	**PO**	**FO**	**FID**	**NC**	
Courtois	X	X	X												Webcrawler
Scoville			X												Excite
Kimmel			X	X											*Lycos*
Davis	X				X										*Alta Vista*
Westera	X														*Alta Vista*
C | Net	X		X				X								*Alta Vista*
Slot		X	X												*Alta Vista*
Zorn	X			X	X						X	X			*Alta Vista*
Lebedev					X			X							*Alta Vista*
Peterson					X							X			Hot Bot
Maldonado	X				X										*Alta Vista*
Notess (a)					X										*Google*
Notess (b)						X									*Google*
Sullivan (a)					X										*Google*
Júpiter Med									X						MSN
Nielsen									X						MSN
Statmarket										X					*Alta Vista*
Sullivan (d)														X	*Google*
Winsip	X		X		X						X	X			*Lycos*
Stobart													X		*Alta Vista*

PB: posibilidades de búsqueda. **TR**: tiempo de respuesta. **GUI**: Interface gráfica de usuario. **DOC**: documentación. **TI**: tamaño del índice. **CRE**: crecimiento porcentual tamaño índice **ACI**: acierto. **FA**: frecuencia de actualización. **AUD**: audiencia. **TRA**: tráfico redirigido. **PO**: porción de página indexada. **FO**: formato de presentación y de alineamiento. **FID**: fidelidad. **NC**: número de consultas que recibe el motor.

Tabla 4.6. Sinopsis de estudios explícitos de evaluación de los motores de búsqueda. Fuentes: elaboración propia.

Estudios experimentales.

Dentro de este grupo figura el conjunto de estudios que han sometido a los distintos motores de búsqueda evaluados, a una serie de ensayos que van más allá de la mera descripción de las características externas de cada motor. Estos estudios han evolucionado con el paso del tiempo y han ganado en complejidad a medida que la web evolucionaba y los problemas de medida subyacentes crecían también.

Revisión de estudios realizados.

A continuación se presentan los estudios más significativos que, dentro de este campo, han sido desarrollados en los últimos cinco años. El propósito de esta revisión no es otro que mostrar cómo han evolucionado estos procesos de análisis y cómo se han ido incorporando medidas cada vez más complejas, que van adaptándose a las características del contexto en el que trabajamos, la web.

Chu y Rosenthal (1996).

El punto de partida de esta revisión, no puede ser otro que el trabajo, ya citado en el apartado anterior, elaborado por Chu y Rosenthal, quienes comentaban que, a su juicio, las diferencias de opinión de los estudios explícitos residían en la ausencia de una metodología clara para la evaluación y se preguntaban "si podíamos llegar a pensar en desarrollar esa metodología que facilite unos resultados que ayuden al usuario a discernir qué herramienta es la más apropiada para sus necesidades específicas" [CHU, 1996].

Para la realización de su propuesta, los autores retoman las seis medidas típicas de evaluación de los SRI propuestas por Cleverdon[75] (cobertura, *exhaustividad*, *precisión*, tiempo de respuesta, esfuerzo del usuario y formato de presentación de los documentos), e instan por una adaptación de las mismas para el nuevo entorno de la web. Ambos autores afirman que estos

[75] Realmente, Chu y Rosenthal citan la primera edición del libro de Lancaster de 1973 (*Information Retrieval Online*), pero ya hemos indicado al principio del anterior capítulo que el establecimiento de este conjunto de parámetros corresponde a Cleverdon, algunos años antes, durante la realización del primer test de Cranfield.

criterios siguen siendo plenamente válidos casi cuatro décadas después de enunciarse, aunque, por ejemplo, no se puede comparar el esfuerzo realizado por un usuario de los años sesenta o setenta para llevar a cabo una recuperación de información en línea con el esfuerzo de cualquiera de nosotros en la actualidad al realizar una búsqueda en la web.

Por las mismas razones, tampoco admiten mucha comparación los formatos de presentación de los documentos recuperados. Los autores plantean un método de evaluación inspirado en la serie de criterios anteriores y que esté adaptado al contexto de la web, donde se valoran cuatro criterios:

1. Composición de los índices: la Cobertura, la Frecuencia de actualización y la Porción de página indexada.
2. Capacidades de búsqueda: las prestaciones que los distintos motores poseen a la hora de recuperar información y la serie de operadores de búsqueda disponibles.
3. Ejecución de la recuperación de información: la *Precisión*, la *Exhaustividad* y el Tiempo de respuesta, procedentes todos ellos de las propuestas clásicas de Lancaster y Fayen [LAN, 1973].
4. Esfuerzo del usuario: la Documentación y la Interface del sistema [CHU, 1996].

Esta propuesta constituye un sustancial salto cualitativo frente a las evaluaciones explícitas anteriormente presentadas, aprovechando de aquéllas los parámetros más fiables e incorporando una serie de cuantificaciones que proporcionan una mejor ponderación de la efectividad de los SRI en la web.

Otro aspecto muy importante de este estudio es su fecha de realización, apenas dos años después de la primera proliferación de sitios web y casi contemporáneo al desarrollo de los primeros motores de búsqueda, lo que le confiere un carácter ciertamente embrionario dentro de una línea seguida a continuación por otros muchos autores. El único aspecto negativo a señalar de este estudio es el escaso número de motores analizados: *Alta Vista, Lycos* y *Excite*, por lo que hace falta recurrir a otras experimentaciones para alcanzar una idea más completa sobre el estado de la cuestión[76].

[76] Aunque quizá también deberíamos ser algo comprensivos ante este escaso número de motores evaluados, en tanto que el número de motores extendidos en aquel momento tampoco era muy elevado.

El análisis del primer criterio propuesto por Chu y Rosenthal no encontró diferencias sustanciales entre los motores, el segundo arrojó unos mejores valores para *Alta Vista* (en torno al 78% sobre una media del 59%), el tercero se decanta claramente hacia *Alta Vista* con la detección de fallos injustificables en *Lycos* (destaca la imposibilidad, en ese momento, de establecer "búsquedas literales"), y en el cuarto criterio, *Lycos* ofrece mejores prestaciones seguido *de Alta Vista*.

Leighton y Srivastava (1995-1999).

Otros autores de similar importancia a los anteriores, son Leighton y Srivastava, quienes han venido publicando diversos trabajos desde hace seis años. En el primero de ellos, Leighton, evaluaba cuatro motores: *Infoseek*, *Lycos*, *Webcrawler* y *WWWWorm*. Tras la formulación y ejecución de ocho preguntas, calculaba una serie de medidas basadas en la *relevancia*, considerando además la ratio de *acierto único* (documentos recuperados sólo en un motor), y el número de enlaces erróneos junto al número de documentos duplicados. El resultado final de este estudio destaca a *Lycos* e *Infoseek* sobre los otros dos [LEI, 1995].

Resultan particularmente curiosas las primeras frases del autor, quien, al principio de la página web donde se recoge una copia de este artículo, indica textualmente: "me encuentro sorprendido con la popularidad de este estudio". Sorprendido porque los datos incluidos en el mismo tenían un corto período de vida y estoy seguro de que los resultados se encuentran absolutamente obsoletos" [LEI, 1995].

Quizás sea esta afirmación la que le lleva a realizar otros trabajos en el mismo campo. Dos años más tarde, contando ya con la colaboración de Srivastava, procede a la evaluación de cinco motores de búsqueda: *Alta Vista*, *Excite*, *Hotbot*, *Infoseek*, y *Lycos*[77]. El período de evaluación abarcaba de enero a marzo de 1997 y en el mismo, los autores introdujeron una interesante variante a la manera típica de calcular la *precisión,* diseñando una función de evaluación que confiere un peso específico a la capacidad de colocar documentos relevantes dentro de los primeros veinte entregados como respuesta al usuario. Esta función, denominada "*First 20 Precision*"

[77] La primera parte de este estudio aparece publicada en la página web <http://www.winona.msus.edu/library/webind2/webind2.htm> y nosotros la hemos recogido en el apartado de referencias con la indicación [LEI, 1997]; aunque la publicación definitiva de los datos aportados por este interesante trabajo aparece en 1999 en la revista *Journal of American Society for Information Science* [LEI, 1999].

mide, al mismo tiempo, la *precisión* y el acierto de mostrar los documentos relevantes antes que los documentos que no son relevantes. En este estudio se penaliza la existencia de enlaces inactivos, de manera que aquellas páginas que no hayan sido actualizadas hace bastante tiempo influirán de forma decisiva en los resultados [LEI, 1997].

Los resultados obtenidos en el estudio de la *precisión* destacan a *Alta Vista*, *Excite* e *Infoseek* (por este orden). Cuando se considera el alineamiento, se aprecian pocas diferencias pero cambia el orden: *Infoseek*, *Excite* y *Alta Vista* [LEI, 1999].

Realmente, lo más importante de este trabajo no son los resultados de *precisión* mostrados por los motores, lo verdaderamente interesante es la aportación de su función de *precisión*, la cual posibilita una aproximación desde una medida de efectividad basada en la *relevancia* hacia el conjunto de medidas orientadas al usuario[78]. Esta función de Leighton y Srivastava favorece a aquellos motores que presentan los documentos relevantes antes que los no relevantes y, en la práctica, esta orientación hacia el usuario se encuentra plenamente justificada porque, generalmente, no resulta de igual interés un resultado devuelto entre los cinco primeros documentos que otro que aparezca en el lugar vigésimo cuarto (por ejemplo), y además se corre el peligro de que el usuario no llegue tan lejos en su lectura.

Arents (1995-2001).

Algunos estudios explícitos ha evolucionado en estudios experimentales en sucesivas repeticiones. Un ejemplo de esta evolución es el desarrollado por Arents desde 1995. En este estudio se seleccionan distintos motores de búsqueda y se los clasifica según una escala: {Mejor - Muy bueno - Bueno – Útil}. El autor basa sus apreciaciones en la facilidad de uso y en la efectividad del motor, entendida ésta como la cantidad, *precisión* y legibilidad de los documentos recuperados. Actualmente, esta clasificación presenta a los motores *Google*, *Alta Vista*, *All the Web*, *Northern Light* y *Raging* (prototipo de motor de búsqueda que está desarrollando *Alta Vista*), como los mejores. Resulta significativo comparar esta lista con la de hace un año aproximadamente[79]: *Northern Light*, *Google*, *Alta Vista* y *Excite*.

---

78 Este conjunto de medidas ha sido presentado en el capítulo anterior.

79 Martínez Méndez, F.J. p. 5

Bharat y Broder (1997).

Este estudio pretende medir "los tamaños relativos de los índices de los motores de búsqueda analizados y su grado de solapamiento, de manera que, para cada par de motores E1 y E2 se calculaba su tamaño relativo (la ratio Tamaño (E1)/Tamaño (E2)) y la fracción del índice del motor E1 que se encuentra indexada por el motor E2, expresada en un porcentaje del tamaño de E1" [BHA, 1998a]. Los motores analizados fueron cuatro: *Alta Vista*, *Excite*, Hotbot e Infoseek, y se tomaron datos en dos períodos de tiempo, junio y noviembre de 1997.

Con relación al solapamiento, los cambios producidos en este período de seis meses fueron insignificantes, de hecho, la única variación apreciable es un aumento de la cobertura de *Excite*. El grado de solapamiento estimado fue del 1.4% del total de los documentos indexados de forma simultánea por los cuatro motores en el mes de noviembre (por un 0.9% obtenido en junio).

Los tamaños relativos estimados presentan cifras consistentes de un experimento a otro, aunque cambia el orden, ya que en junio *Hot Bot* supera a *Alta Vista*, *Excite* e *Infoseek* (por este orden) y en noviembre, *Alta Vista* supera a Hot Bot. En el experimento de junio, *Hot Bot* indexaba el 47% de todos los documentos recuperados por los cuatro motores y, en noviembre, *Alta Vista* indexaba un 62% de los documentos recuperados. Seis meses más tarde, Bharat y Broder actualizan los datos de este estudio [BHA, 1998b], obteniendo porcentajes similares. En este segundo estudio calculaban, en marzo de 1998, que la web crecía a un ritmo de *veinte millones de páginas mensuales*, lo que suponía que duplicaría el tamaño de noviembre de 1997 (aproximadamente 200 millones de páginas) en nueve meses, hacia finales del año 1998.

Ralph (1997).

Este trabajo surge como producto de unos análisis preliminares realizados en 1996, en el seno de un curso organizado por el *Departamento de Información y Documentación de la Universidad de Carolina del Norte en Greensboro*. En el mismo se analiza un conjunto de siete motores de búsqueda: *Alta Vista*, *Excite*, *Hot Bot*, *Infoseek*, *Lycos*, *Open Text*, *Webcrawler* y un directorio: *Yahoo*.

Sobre estos sistemas se realizaron veinticinco preguntas, la mayor parte de ellas correspondían a términos individuales, nombres de personas o

instituciones o materias. El estudio analizó el número de documentos recuperados por cada motor, la *precisión* (tomando como muestra para su determinación los veinticinco primeros documentos devueltos para cada cuestión), la *exhaustividad* relativa y el alineamiento. Para estudiar este último aspecto, establecieron unos criterios de clasificación que valoraba la presencia de los documentos relevantes al principio de la secuencia de documentos entregados como resultado.

Los resultados de la *precisión* obtenidos en este estudio dividen a los motores analizados en dos grupos claramente diferenciados. El primero de ellos agrupa a *Alta Vista*, *Excite*, Hot Bot e Infoseek (que alcanzan porcentajes superiores al 60%), frente al resto de motores cuyos valores oscilan entre el 30% de *Lycos* y el 35% de *Webcrawler*. El motor que mejor porcentaje alcanza de *exhaustividad* relativa es Hot Bot, que supera ligeramente el 25%. La ponderación del alineamiento de los documentos igualaba a Hot Bot y *Excite*, aunque los autores se inclinan por considerar mejor al primero de estos motores porque devuelve casi el doble de documentos que segundo.

Wishard (1998).

Wishard lleva a cabo un estudio de la *precisión* de diecisiete motores de búsqueda, realizando sobre ellos una serie de preguntas relacionadas con la Geología y no encuentra diferencias significativas entre todos ellos. En sus cálculos sobre la *precisión* obtenida por cada motor, Wishard considera que *Infoseek* es ligeramente superior a *Lycos* y *Alta Vista*, aunque él mismo reconoce que sus resultados no son tan concluyentes como para contradecir oros trabajos anteriores que presentan diferentes conclusiones (menciona, entre otros, los trabajos de Leighton y Srivastava).

Lawrence y Giles (1998).

El análisis de la cobertura de los motores de búsqueda centra el principal interés de este estudio. Sus autores estiman también el tamaño de los distintos índices de los motores, su grado de solapamiento y el porcentaje de enlaces erróneos que proporciona cada motor. El conjunto de seis motores evaluados es: *Alta Vista*, *Excite*, *Hot Bot*, *Infoseek*, *Lycos* y *Northern Light*. En su estudio, los autores comentan una idea muy extendida entre los usuarios de Internet: "estos sistemas indizan más o menos los mismos documentos y recopilan un amplio porcentaje de la totalidad de la web" [LAW, 1998].

A partir de una serie de 575 cuestiones realizadas por empleados de los laboratorios de investigación de *NEC Corporation*, se creó una lista con la

totalidad de documentos recuperados que cumplían dos condiciones: que hubieran sido localizados por todos los motores analizados y que incluyeran en su texto completo los términos por los que habían sido recuperados, es decir, que su presencia en la respuesta se debiera a la presencia de todas las palabras que hubieran sido empleadas en la búsqueda[80]. Los documentos duplicados fueron eliminados de la relación de documentos devueltos por cada motor y se analizaron los primeros seiscientos documentos de cada motor. Como resultado de este análisis, se deduce que el motor *Hot Bot*, con un 57.5%, es el motor que más documentos comunes posee en su relación, seguido de *Alta Vista* (con un 46.5%). En los cálculos de estimación del tamaño de la web, que se llevan a cabo a partir de los datos anteriores de solapamiento, se tasa en un 34% el porcentaje de web indexado por *Hot Bot*. La unión del índice de este motor y del índice de *Alta Vista*, produciría un conjunto de más de 320 millones de páginas, mientras que la unión de *Infoseek* y *Lycos* produciría un índice de 90 millones de páginas y la de *Excite* y *Northern Light* alcanzaría los 230 millones.

Estos valores estimados por Lawrence y Giles poseían mucha importancia en su momento, en tanto que "otras evaluaciones realizadas con anterioridad, tales como la de Forrester (quien estimaba el tamaño de la web en 75 millones de documentos) o la del ejecutivo jefe y creador de *Alta Vista*, Louis Monier (quien estimaba que la web contenía entre 100 y 150 millones de páginas), han subestimado el tamaño de la web" [LAW, 1998].

En el análisis del número de enlaces erróneos facilitados por cada motor, *Hot Bot* destaca por ser el que más proporciona (5 %) frente a *Lycos* (1.6%), es decir, al menos en este estudio, el índice de tamaño más pequeño proporciona menos errores entre sus enlaces que el de mayor tamaño.

Gordon y Pathak (1999).

Gordon y Pathak desarrollan un estudio mucho más amplio que el desarrollado por Chu y Rosenthal, ya que analizan hasta siete motores de búsqueda: *Hot Bot*, *Magellan*, *Infoseek*, *Excite*, *Lycos*, *Alta Vista* y *Open Text*; junto a un directorio, *Yahoo* [GOR, 1999].

---

80 No hay que olvidar, tal como mencionan los autores de este estudio, que los motores de búsqueda devuelven muchas veces páginas donde no aparecen todos los términos solicitados, especialmente si el evaluador no ha exigido esa presencia en la redacción de la búsqueda o el sistema no se lo permite en su interface de usuario.

Estos autores consideran como muy importante la naturaleza intrínseca de los motores de búsqueda, especialmente todo lo relacionado con las fuentes de información a las que acceden. Estos ingenios indexan páginas (bien por su cuenta o bien a sugerencia de sus usuarios), de una temática muy heterogénea y sin establecer restricciones al contenido y al formato de esos documentos, excepto a aquellas páginas cuyo contenido es presuntamente delictivo. Esto viene a colación en tanto que, para Gordon y Pathak, no podemos extrapolar los tests desarrollados sobre los SRI tradicionales a este nuevo entorno, al tener los SRI tradicionales unas bases de conocimiento mucho más estructuradas y una naturaleza informativa más homogénea que en el contexto de la web, siempre más abierto tanto en estructura de los documentos como en la naturaleza de los contenidos.

Por último, recuerdan que a la hora de expresar la necesidad de información del usuario en los términos empleados por el motor de búsqueda, cada uno de ellos introduce una serie de limitaciones y además, hay que considerar que cada motor de búsqueda emplea un algoritmo distinto, que el contenido de los índices gestionados por cada uno de ellos es diferente y que, incluso si varios motores de búsqueda llevaran a cabo la misma operación de recuperación de información sobre el mismo conjunto de documentos, el resultado sería diferente en tanto que cada motor de búsqueda hace uso de un algoritmo de alineamiento distinto.

Es por todo ello, por lo que Gordon y Pathak afirman que hace falta una serie de características más específicas que complementen a los parámetros tradicionales empleados en la evaluación de la recuperación de información. Estos autores proponen un total de siete:

1. Hay que tener en cuenta que los experimentos que puedan llevarse a cabo sobre estos motores de búsqueda, suelen realizarse por medio de usuarios experimentados en la recuperación de información, quienes, en virtud de su conocimiento de los sistemas pueden introducir algunas modificaciones sobre las necesidades de información originales e incluso pueden llegar a favorecer a un motor frente a otro en virtud de la costumbre de uso del mismo. Este pequeño sesgo, no debe obviarse, ya que puede atenuar la frescura derivada de las interrogaciones que llevan a cabo los usuarios individuales de estos sistemas.
2. Si en una operación de recuperación de información, una persona busca documentos que alguien más ha encontrado previamente, la necesidad de información de esa persona debe ser capturada tan en

el contexto como sea posible, sin que incidan búsquedas anteriores o paralelas.

3. Parece un hecho suficientemente constatado que se diseña un número suficientemente grande de búsquedas con la idea de producir evaluaciones satisfactorias, es decir, se encuentran inducidas al éxito de antemano.
4. Los estudios realizados con detenimiento han de incluir más motores de búsqueda.
5. La efectividad de los motores de búsqueda debe ser analizada explotando las características especiales de cada uno de ellos.
6. Los juicios de *relevancia* deben realizarse por los individuos que precisan de la información, alejándonos de los juicios de expertos.
7. Los experimentos bien dirigidos resultan necesarios para obtener unas medidas significativas de la realización de las búsquedas. Deben seguir un diseño experimental apropiado, convirtiendo en aleatorio el orden en el que los documentos se presentan a los evaluadores para superar los problemas vinculados al orden de presentación y adaptándose a medidas tradicionales (como la *precisión* y la *exhaustividad*), permitiendo que los resultados se evalúen dentro de un contexto familiar y empleando tests estadísticos para medir las diferencias detectadas entre los distintos motores de búsqueda [GOR, 1999].

En su experimento, Gordon y Pathak contaron con la ayuda de treinta y seis profesores familiarizados con el uso de los distintos motores de búsqueda evaluados. Estos profesores analizaron las distintas necesidades de información sobre Economía y Nuevas Tecnologías que les transmitieron unos usuarios por medio de unos cuestionarios y efectuaron sus búsquedas en todos los sistemas, indicando el éxito o fracaso de los documentos recuperados entre las veinte primeras respuestas que proporcionaba el motor de búsqueda. El análisis posterior se centró en examinar el comportamiento de la *exhaustividad* y *precisión* en cada cuarteto de cinco documentos recuperados[81], ofreciendo un importante número de conclusiones, entre las que destacamos:

[81] Se establecieron cuatro grupos de cinco documentos agrupados conforme a su orden en la recuperación de información (1º-5º, 6º-10º, 11º-15º y 16º-20º).

1. La efectividad media en la recuperación de información es baja, alrededor del diez por ciento en los primeros diez documentos recuperados.
2. Se detectan diferencias estadísticas considerables entre la *precisión* de los motores de búsqueda analizados. *Alta Vista*, *Open Text* y *Lycos* resultan los mejores y *Yahoo* ocupa el último lugar. Otra cuestión importante a destacar es la amplia distancia entre el primero, alrededor del 40%, y el segundo, sobre el 20%.
3. En el caso de la *exhaustividad* no se detectan diferencias sustanciales entre los motores.
4. La realización de una búsqueda aparece más fuertemente vinculada al algoritmo de localización de la información que a los tipos de búsqueda que permite.
5. Aunque los motores siguen, en líneas generales, los principios generales de la recuperación de información, su funcionamiento varía de alguna manera ya que es clara la tendencia a recuperar documentos irrelevantes o lejanamente relevantes de algunos de ellos, hecho difícilmente producible en el caso de un sistema de recuperación de información tradicional.

Gordon y Pathak también aportan otras conclusiones no menos importantes, una de ellas se refiere a que los índices de los motores poseen tamaños muy diferentes (algunos son diez veces más grandes que otros) y, aunque algún motor, en su publicidad suela afirmar lo contrario, ninguno pretende realmente indexar toda la web. Los motores también difieren en la actualización periódica de los datos, en la posibilidad de que los usuarios añadan páginas por su cuenta, en el plazo de tiempo que se tarda en incorporar una nueva página indexada tras tener noticia de su existencia y en el seguimiento de la disponibilidad de los enlaces.

Otro apartado interesante de este estudio es el dedicado al cálculo del grado de solapamiento. Como cada motor de búsqueda sigue distintos criterios a la hora de recopilar páginas de la web y actualizar sus índices, es interesante conocer qué porción de la web poseen en común los diversos motores de búsqueda. Una parte de los documentos en común se debe, seguramente al interés de los administradores de estos sitios web en incorporarlos (de forma manual y voluntaria) a los índices de los motores de búsqueda; aunque la política que siguen los motores de búsqueda para admitir a estas páginas sugeridas no siempre es la misma, y no todas llegan a formar parte de sus índices.

De otro lado, es lógico suponer que una misma página pueda ser visitada por los agentes que recopilan información para los motores de búsqueda. A pesar del ingente tamaño de la web, algunos de estos sistemas son capaces de acceder a más de cinco millones de páginas en un día y la probabilidad de que una página sea visitada por al menos dos motores distintos es de suponer que no es pequeña. Gordon y Pathak estudian el solapamiento detectado en su estudio, bajo dos perspectivas:

a) El grado de solapamiento existente entre los documentos recuperados por los ocho motores analizados
b) El grado de solapamiento existente entre los documentos recuperados por los ocho motores analizados, que han sido considerados relevantes

Resultan sorprendentemente bajos para los autores los resultados del primer tipo de solapamiento, cerca de 150 documentos de los 160 documentos recuperados por los motores analizados en primer lugar (los 20 primeros de cada uno de los 8 motores), han sido recuperados únicamente por un solo motor. Este porcentaje (inferior al 10% de solapamiento) se mantiene constante cuando el análisis se prolonga a los 50, 100 y 200 primeros documentos recuperados por cada motor.

El segundo tipo de solapamiento presenta unos valores mucho mayores de coincidencia que el anterior, ya que el número de documentos relevantes recuperados por un único motor oscila entre el 22.45% y el 25.4%, es decir que el grado de solapamiento no desciende nunca del 70%. Esta segunda serie de datos resulta de mucho interés, ya que permiten suponer que los motores de búsqueda muestran muchos documentos distintos unos de otros pero, en principio, un porcentaje importante de los relevantes serán parte de la respuesta de cada uno de ellos. Aunque esta afirmación se encuentra sujeta a muchos condicionantes y debe ajustarse aún mucho más, permite una aproximación a la solución de la cuestión planteada al principio de este capítulo, sobre cuál era el mejor motor de búsqueda, ya que, los datos de Gordon y Pathak desprenden que una porción significativa de los documentos relevantes forma parte de los índices de, al menos dos, motores de búsqueda.

Gwizdka, J. and Chignell (1999).

En la misma línea que el estudio anterior, estos autores realizan un estudio de las publicaciones anteriores sobre este campo, aunque, a diferencia de los anteriores, resaltan estudios donde se considera interesante la capacidad de los enlaces de una página para determinar el grado de *relevancia* de la

misma. Otra cuestión que ponen en duda estos autores, es la tradicional asignación binaria de los valores de la *relevancia* (documento relevante – documento no relevante), que difícilmente permiten recoger la totalidad de los casos. Ellos propugnan por la introducción de algunos valores intermedios en esa escala.

En su estudio, estos autores analizan un importante número de aspectos: la *precisión* (entendida en los términos de la función definida por Leighton y Srivastava), el orden de presentación de los documentos o alineamiento (estableciendo una función diferencial para la *precisión* de los documentos conforme aparezcan alineados), el esfuerzo del usuario, la longitud esperada de búsqueda, el número de enlaces erróneos y el número de enlaces duplicados.

El análisis abarca a tres motores de búsqueda (*Alta Vista*, *Hot Bot* e *Infoseek*), en cuatro dominios de países distintos (Alemania, Austria, Polonia y Reino Unido), y a los tradicionales dominios .com (empresas con ánimo de lucro) y .org (sitios corporativos o de instituciones sin ánimo de lucro). La función diferencial de la *precisión*, que refleja cómo de bien han sido alineados los primeros veinte documentos, tras el experimento indica que *Infoseek* tiende a concentrar los documentos más relevantes dentro de los diez primeros recuperados, aunque, por el contrario, en algunos casos (13 de 24), este motor ha devuelto un número menor de 20 documentos, situación que ha favorecido a este motor para alcanzar buenos valores con esta medida. Los otros dos motores no han presentado diferencias significativas en este diferencial.

Como resultados globales de este experimento, *Alta Vista* presenta mejores resultados que los otros dos motores, en términos de *precisión* y de diferencial de *precisión*[82].

El solapamiento entre los documentos es bajo, fundamentalmente porque "emplean diferentes procedimientos de localización de los documentos en la consulta y porque sus métodos de recopilación e indización de los documentos son sustancialmente distintos" [GWI. 1999]. El número de aciertos únicos detectados es alto, debido al escaso grado de solapamiento y a que

82 Los autores comparan los resultados de sus cálculos de *precisión* e indican que son consistentes con las conclusiones aportadas por Chu y Rosenthal y por Leighton y Srivastava, tal como hemos presentado anteriormente.

el estudio analiza un número relativamente moderado de respuestas de cada motor[83].

El experimento no detecta efectos significativos de la *precisión* entre los motores de búsqueda y los dominios analizados, siendo *Alta Vista* el que presenta mejores índices de cobertura en todos ellos[84], de hecho, los autores indican la escasísima cobertura del motor *Infoseek* fuera del dominio .com y su poca presencia en los dominios de lengua no inglesa (90.000 páginas frente a un millón de *Hot Bot*, por ejemplo, en las fechas de realización del experimento).

Incluso el motor *Hot Bot*, a pesar de poseer una mayor presencia en dominios de lengua no inglesa, no puede compararse con uno como *Alta Vista*[85], que indexa muchas más páginas y que parece ser más insensible a los cambios de lengua y de dominio.

BIBLIOTECA VIRTUAL

Ming (2000).

Similar línea de trabajo sigue el completo estudio de Ming, que analiza diversos parámetros, tales como la *precisión*, el tiempo de respuesta del sistema, la interface de usuario, el número de aciertos e introduce un factor que denomina "sensibilidad", factor sobre el cual incide la calidad de los enlaces devueltos por el motor de búsqueda.

Por desgracia, este estudio sólo afecta a *Yahoo*, *Alta Vista* y *Lycos*, y ofrece como resultados más destacados el hecho de no encontrar diferencias significativas en el número de documentos devuelto por cada sistema; prefiere el interface y el tiempo de respuesta ofrecido por *Yahoo* (aunque reconoce que es una opinión subjetiva).

En lo relacionado con el análisis de la *precisión*, destaca que el valor medio de los tres motores evaluados en los primeros diez documentos devueltos es ligeramente superior al obtenido cuando sólo se toman en cuenta los cinco

---

83 Evidentemente, si en el experimento se analizaran, por ejemplo, las doscientas primeras referencias devueltas por cada motor, el número de aciertos únicos descendería en comparación a cuando se leen sólo las diez o veinte primeras referencias.

84 Este dato puede confirmarse con los resultados del estudio (anteriormente expuesto), Notess, quien, en sus estimaciones de tamaño del índice de los motores de búsqueda, sitúa a *Alta Vista* por delante de Hot Bot y de Infoseek.

85 Esta reflexión de los autores resulta particularmente importante a la hora de establecer las bases para la realización de un experimento, ya que, si el número de documentos indexados tiene alguna incidencia sobre las medidas a calcular, algunos motores de búsqueda van a encontrarse en inferioridad de condiciones.

primeros documentos devueltos. En el primer caso, *Yahoo* supera a *Lycos* y *Alta Vista* (por este orden), y en el segundo caso se mantiene similar clasificación. En relación al parámetro de la "sensibilidad", Ming afirma que *Yahoo* supera a los otros dos, resaltando los pobres valores de *Alta Vista* [MIN, 2000].

Notess (2000).

En el apartado de los estudios explícitos se ha hecho mención a los diversos análisis realizados por Notess, recogidos en la web *Searchengine Showdown*. Dentro del campo de los estudios experimentales, interesan los cálculos de solapamiento, acierto único y enlaces fallidos que fueron realizados en febrero de 2000, algunos meses más tardes del trabajo de Gordon y Pathak [NOT, 2000].

De un total de 298 documentos diferentes recuperados por cuarenta motores, 110 fueron devueltos únicamente por un solo motor (casi el 37%), y el porcentaje de documentos que al menos aparecía en dos motores alcanzaba el 26,51%. Estos valores presentan unos porcentajes algo mayores que los reflejados en el estudio anterior, influyendo en este aumento el crecimiento que han experimentado los índices de los motores y también el alto número de motores analizados.

A continuación se analizaron los 110 documentos devueltos únicamente por un motor, con la idea de determinar la distribución de estos documentos entre los motores. El resultado fue: *Fast* (ahora *All the Web*) con un 40% de ocurrencias, *Alta Vista* (23%), *Excite* (18%), *Northern Light* (10%) y *Google* (6%). En este punto resulta interesante resaltar los cambios que este estudio sufre con el tiempo y con la evolución del índice de los motores, ya que un estudio realizado unos cuatro meses antes, el motor con mayor porcentaje de aciertos únicos era *Northern Light* con un 30%.

En último lugar, Notess analizó el porcentaje de enlaces fallidos que proporcionan los distintos motores de búsqueda. El estudio estaba basado en el análisis de los cien primeros documentos devueltos en tres búsquedas por cada motor.

Notess calcula dos tipos de enlaces fallidos, el primero de ellos engloba a todos los errores "tipo 400" del protocolo HTTP (página no encontrada, acceso denegado, acceso prohibido) y errores que se producen en la conexión; el segundo sólo refleja los enlaces fallidos. La Tabla 4.7 resume el resultado del estudio:

	% Enlaces fallidos	% Errores 400 sólo
AltaVista	13.7%	9.3%
Excite	8.7%	5.7%
Northern Light	5.7%	2.0%
Google!	4.3%	3.3%
HotBot	2.3%	2.0%
Fast	2.3%	1.8%
MSN Inktomi	1.7%	1.0%
Anzwers	1.3%	0.7%

Tabla 4.7. Porcentajes de fallos en las referencias ofrecidas por los motores de búsqueda, Fuente: Notess, G. R. Search Engine Statistics [En línea]. Bozeman, MT: Notes.com, 2000. <http://www.searchengineshowdown.com/stats/> [Consulta: 10 noviembre 2001]

Ljoslund (1999-2000).

De los trabajos realizados en los últimos años, destacan especialmente los llevados a cabo por la investigadora noruega Ljoslund, de quien resulta conveniente resaltar dos aportaciones. La primera de ellas se centra en el análisis de las cuestiones más importantes a considerar en la evaluación de los motores de búsqueda, la segunda se centra más en el análisis comparativo del funcionamiento de un importante número de motores de búsqueda cuando se les interroga por palabras de uso poco corriente[86]. El primero de los trabajos de Lojslund, presentado en la *Conferencia SIGIR'99*[87], analiza la *relevancia* de los documentos recuperados, introduciendo una modificación en su cálculo con respecto a la mayoría de los estudios anteriores: considera un nivel intermedio de *relevancia*. Es decir, rompe la tradicional dicotomía de documento relevante – documento no relevante, introduciendo el concepto de documento parcialmente relevante. Trasladando estos valores a una escala decimal, Lojslund establece tres posibles valores para la *relevancia* de un documento: 0, 0.5 y 1 [LOJ, 1999].

Otro aspecto muy importante de este estudio es su proximidad al problema analizado, es decir, al contexto de la web, acercamiento no del todo detectado en los autores anteriores. Lojslund se plantea, antes de comenzar a analizar la *relevancia* de los documentos recuperados, cómo reaccionar

[86] De hecho, la autora clasifica a esta serie de términos como "palabras raras".

[87] SIGIR'99: acrónimo de "22*th International Conference of Research and Development in Information Retrieval*".

ante situaciones muy frecuentes en la web, por ejemplo: un motor devuelve dos páginas exactamente iguales en contenido pero que se encuentran alojadas en sitios web distintos[88]. La autora se pregunta si son relevantes ambas páginas. Los contenidos seguramente sí, pero la segunda página no va a ser de ningún interés para el usuario, así que su *relevancia*, para Lojslund es nula.

Esta anterior decisión, aunque ciertamente susceptible de discusión y de posibles reinterpretaciones, no puede pasar desapercibida, en tanto que el análisis de la *relevancia* que se lleva a cabo en este estudio se encuentra estrechamente vinculado al contexto en el que se desarrolla: la web. Fundamentalmente, este aspecto del estudio de Lojslund le confiere una mayor representatividad frente a otros estudios que simplemente trasladan los criterios empleados en los estudios realizados sobre los tradicionales SRI a la web si entrar a considerar situaciones de nuevo signo.

La autora compara únicamente tres motores: *Alta Vista*, *Google* y *All the Web*. El primero es un motor de consolidada posición y los otros dos son nuevos proyectos que buscan hacerse un hueco dentro de este amplio conjunto[89]. El experimento muestra que, cuando no se considera la *relevancia* parcial, se obtiene una *precisión* media de 0.4 para *Alta Vista*, 0.7 para *Google* y de 0.4 para *All the Web*.

Si se considera la *relevancia* parcial, suben un poco estos valores: 0.5 para *Alta Vista*, 0.9 para *Google* y 0.5 para *All the Web*. Otro dato importante medido es la posibilidad de encontrar un documento relevante en el primer lugar de la lista de documentos devueltos por cada motor, siendo de nuevo a *Google* el motor que más destaca (80% frente al 50% del *All the Web* y el 30% obtenido por *Alta Vista*).

El segundo trabajo de Losjslund analiza la efectividad de la recuperación de información de veinte sistemas entre los que mezcla motores, directorios y metabuscadores. A todos les pide que localicen documentos con palabras de uso poco frecuente. El total de las palabras solicitadas es diez[90].

---

88 Esta situación es lo que se denomina "mirror site". No es una situación extraña, especialmente cuando los motores de búsqueda devuelven páginas de directorios (*Yahoo*, por ejemplo), que cuentan con réplicas o "mirrors" en diversos países. El contenido es idéntico pero la URL es distinta.

89 El estudio está publicado en 1999, cuando *Google* y *All the Web* iniciaban su actividad.

90 Palabras empleadas en el experimento: "haliography, pogonotomy, bivious, unharmless, peleidou, apodyopsis, materteral, fluctisonant, curmedgeon, tameins".

	Docs. Recuperados	Docs. Relevantes con duplicados	Docs. Relevantes	Búsquedas no vacías
Fast	93	74	70	9
AskJeeves	86	56	49	10
NorthernLight	81	44	38	9
InferenceFind	69	53	41	9
EgoSurf	66	50	41	7
Excite	66	47	35	9
Oingo	55	43	35	9
AltaVista	50	45	38	9
msn	47	37	27	8
Yahoo!	40	34	28	8
InfoTiger	39	36	29	9
Google	38	33	27	6
snap.com	38	30	22	8
HotBot	29	23	19	7
DirectHit	27	5	5	3
Infoseek	18	13	11	6
Lycos	12	11	10	3
GoTo.com	9	6	6	2
Euroseek	6	6	6	4
Webcrawler	5	5	4	5

Tabla 4.8. Resultados del estudio de Ljoslund de búsqueda de palabras raras. Fuente: Lojslund, M. Evaluation of twenty Web search engines on ten rare words ranking algorithms. [En línea]. Trondheim: Sør-Trøndelag University, 2000. <http://www.aitel.hist.no/~mildrid/dring/paper/Comp20.doc> [Consulta: 21 de noviembre de 2001]

La Tabla 4.8 recoge los documentos recuperados por cada SRI y cuántos son relevantes (sin discernir los documentos duplicados), la tercera columna elimina esos documentos duplicados y la cuarta columna muestra el número de búsquedas que no han resultado vacías, es decir, que han devuelto al menos un documento (este valor resulta especialmente interesante en este estudio porque las palabras solicitadas son de uso poco frecuente). Lojslund no emplea aquí la "*relevancia* parcial" [LOJ, 2000]. Los resultados muestran que tanto *Fast* como *AskJeevesm,* devuelven una mayor cantidad de documentos y poseen mejor *precisión*. La Tabla 4.9 muestra los resultados ordenados según la *precisión*. De su análisis se deduce que la *precisión* es mejor para los motores que poseen escasa cobertura y que pocas veces presentan listas vacías de documentos (es decir, que no devuelven nada).

	Cobertura	***Precisión***	***Exhaustividad***	***Exhaustividad única***[91]
Euroseek	0,01	1,00	0,03	0,03
Webcrawler	0,02	1,00	0,03	0,04
Lycos	0,05	0,95	0,06	0,08
InfoTiger	0,21	0,94	0,26	0,35
Yahoo!	0,09	0,93	0,15	0,23
AltaVista	0,20	0,91	0,29	0,40
Excite	0,24	0,90	0,29	0,38
msn	0,10	0,88	0,15	0,18
Google	0,15	0,87	0,21	0,32
Snap.com	0,09	0,85	0,12	0,17
Oingo	0,23	0,84	0,32	0,42
Fast	0,34	0,82	0,42	0,58
InferenceFind	0,28	0,81	0,38	0,46
EgoSurf	0,21	0,80	0,29	0,39
AskJeeves	0,39	0,79	0,43	0,57
GoTo.com	0,02	0,79	0,03	0,04
HotBot	0,07	0,78	0,10	0,15
Infoseek	0,06	0,75	0,08	0,11
NorthernLight	0,24	0,55	0,20	0,29
DirectHit	0,05	0,52	0,02	0,02

Tabla 4.9. Resultados del estudio de Ljoslund de búsqueda de palabras raras ordenados por la *precisión*. Fuente: Lojslund, M. Evaluation of twenty Web search engines on ten rare words ranking algorithms. [En línea]. Trondheim: Sør-Trøndelag University, 2000. <http://www.aitel.hist.no/~mildrid/dring/paper/Comp20.doc> [Consulta: 21/11/2001

La propia autora resume los resultados obtenidos en ambos ensayos en la siguiente Tabla 4.10:

Cobertura	***Precisión***	***Exhaustividad***	***Exhaustividad* única**
AskJeeves(0.39)	Euroseek, *Webcrawler* (1.00)	AskJeeves (0.43)	*Fast*(0.58)
Fast(0.34)		*Fast*(0.42)	AskJeeves(0.57)
InferenceFind(0.28)	*Lycos* (0.95)	InferenceFind (0.38)	InferenceFind(0.46)

Tabla 4.10. Tabla resumen del estudio de Ljoslund de búsqueda de palabras raras ordenados por la *precisión*. Fuente: Lojslund, M. p.6 [Consulta: 21/11/2001]

[91] Número de documentos localizados sólo en un motor de búsqueda dividido por el número total de documentos únicos localizados.

Como promedio, cada documento es localizado 2.8 veces por los distintos motores analizados, 157 documentos (casi el 50%) fueron localizados por un único motor de búsqueda y un documento fue devuelto por 15 motores.
Siguiendo con este cómputo, ocho motores de búsqueda entregaron documentos que ningún otro devolvió como resultado, todos estos datos confirman lo avanzado en otros estudios: que el grado de solapamiento de los motores es pequeño. Si uno desea recuperar muchos documentos diferentes, debe consultar muchos motores. Ljoslund presenta un caso extremo: "incluso los metabuscadores devuelven resultados diferentes. Por ejemplo, casi todos ellos devuelven documentos localizados en el índice de *Alta Vista*, pero *Alta Vista* devuelve por su cuenta documentos únicos que no están recogidos en esos metabuscadores" [LOJ, 2000].
Esta última reflexión, sirve a la autora de preludio para presentar la comparación de los resultados devueltos por los distintos motores por medio de una *función de similitud*, la de Jaccard más concretamente [SAL, 1983] [CHO, 1997]. Con la aplicación de una función de similitud, la autorra pretende averiguar hasta qué punto los conjuntos de documentos devueltos por algunos motores de búsqueda son realmente subconjuntos de los devueltos por otros, tal como puede ser el caso de la mayor parte de los documentos devueltos por el motor Goto.com[92] en este estudio, que están incluidos en la respuesta de *Microsoft Network*, Sanp.com y *Yahoo*. A pesar del bajo grado de solapamiento de los distintos motores evaluados, la función de Jaccard ofrece resultados verdaderamente altos en algunos casos[93], como, por ejemplo, Infotiger y *Excite* (89%), *Microsoft Network* y Sanp.com (84%) y finalmente, *Yahoo* y *Microsoft Network* (80%). Extendiendo el cálculo de la similitud a un conjunto de varios motores (a modo de agrupamiento o "cluster"), los resultados más destacados muestran que el agrupamiento formado por {*AltaVista*, *InferenceFind*, *Oingo*} posee una similitud del 61% y que el formado por {*HotBot*, *Snap.com*, *Yahoo*, *Microsfot Network*} alcanza un valor del 55%.
Más allá de las cifras obtenidas en este segundo estudio y de las conclusiones de la autora, destaca la originalidad del cálculo de la similitud de los resultados devueltos por cada motor, ya que la coincidencia en la composición de los índices de documentos de los distintos motores se ha venido calculando normalmente con base en el porcentaje de solapamiento y de referencia única. Teniendo en cuenta que el estudio de

---

92 Este motor ha cambiado de nombre, ahora es Overture.com <http://www.overture.com> [Consulta: 2 de noviembre de 2001]

93 Se entiende que el valor medio de esta función es el cincuenta por ciento.

Lojslund mide porcentajes de cobertura, *precisión* y *exhaustividad* de los distintos motores evaluados, que el mismo abarca a una considerable cantidad de motores (veinte), y que calcula el grado de solapamiento y de similitud de los resultados, no hay duda alguna sobre que ambos trabajos se tratan de los estudios más completos de los realizados hasta ahora sobre este campo. A pesar de todas esta serie de bondades, los resultados no resultan extrapolables por dos razones: primeramente, en el conjunto de SRI analizados se entremezclan motores de búsqueda, directorios y metabuscadores, y en segundo lugar, las cuestiones planteadas se referían a la presencia de términos de uso poco frecuente en documentos, no a la *relevancia* o no *relevancia* de un documento con un tema objeto de una necesidad de información.

CNET search site olympics (Thomas, 2002).

Thomas, publica en la web de la revista electrónica *C | NET* el resultado de un evaluación de cinco motores de búsqueda (*Alta Vista, Excite, Google, Lycos* y *MSN Search*), realizada a principios del año 2002. En la misma medía dos conceptos, uno experimental y otro más subjetivo (el diseño artístico del sitio web y la presentación de los resultados). El criterio experimental era la *relevancia* de los enlaces devueltos. Los resultados otorgan mejores resultados a *Google* con una amplia ventaja sobre el resto, tal como se refleja en la siguiente tabla:

AltaVista	**Excite**	***Google***	***Lycos***	**MSN Search**
46.40	Descalificado[94]	65.95	49.57	49.08

Tabla 4.11. Tabla resumen del estudio de Sarah Thomas. Fuente: C | NET

Sinopsis de estudios experimentales.

La revisión de los estudios de evaluación de los motores de búsqueda de tipo experimental ha abarcado un total de catorce trabajos, que han empleado, a lo largo de su desarrollo, diecisiete variables diferentes para llevar a cabo sus procesos de análisis. Estos estudios se han sintetizado en la Tabla 4.12.

[94] El motor Excite fue descalificado de la evaluación debido a los cambios que se han producido recientemente en este servicio, que ha pasado a formar parte de otro motor de búsqueda: *Overture*.

	PB	**TR**	**GUI**	**DO**	**TI**	**P**	**E**	**CO**	**FA**	**PO**	**HU**	**EE**	**ED**	**FOR**	**SO**	**LO**	**EN**	
Chu	X	X	X	X		X	X	X	X	X								AVI
Leighton, 1995						X	X				X	X	X					LYC
Leighton, 1999						X												AVI
Arents			X		X	X								X				GOO
Bharat					X										X			AVI
Ralph					X	X	X	X						X				HOT
Wishard						X												LYC
Lawrence					X			X				X			X			HOT
Gordon					X	X	X		X						X			AVI
Gwizdka			X			X						X	X	X		X		AVI
Ming		X	X		X	X											X	AVI
Notess					X			X			X	X			X			FAST
Ljoslund, 1999						X												GOO
Ljoslund, 2000					X	X	X				X				X			FAST
Thomas, 2002						X								X				GOO

PB: posibilidades de búsqueda. **TR**: tiempo de respuesta. **GUI**: Interface gráfica de usuario. **DOC**: documentación. **TI**: tamaño del índice. **P**: *precisión*. **E**: *exhaustividad*. **CO**: cobertura. **FA**: frecuencia de actualización. **PO**: porción de página indexada. **FID**: fidelidad. **HU**: acierto único. **EE**: enlaces erróneos. **ED**: enlaces duplicados. **FOR**: formato de presentación y alineamiento. **SO**: solapamiento **LO**: longitud esperada de búsqueda **AL**: alineamiento **EN**: calidad de los enlaces

AVI: *Alta Vista*. **LYC:** *Lycos*. **GOO:** *Google*. **HOT**: Hot Bot. **FAST:** All the Web.

Tabla 4.12. Sinopsis de estudios experimentales. Fuente: elaboración propia.

El número de estudios realizados es menor que en el caso de los explícitos y se han usado más variables, ocurrencias perfectamente comprensibles ya que un estudio experimental suele ser más complejo en su elaboración. El número medio de variables empleadas por trabajo alcanza el valor de 4.21, resultado de dividir las 59 variables entre los 14 estudios, duplicando así la cifra obtenida en los estudios explícitos. Al igual que en la presentación de los estudios explícitos, en la última columna de la tabla se indica qué motor recibe la mejor valoración en cada estudio. En esta ocasión aparecen citados cinco motores, dos menos que en el caso anterior. Las características

más veces empleadas son: *precisión* (12 veces), tamaño del índice[95] (8 veces), enlaces erróneos y solapamiento (5 veces). La mayoría de los estudios que emplean la *precisión*, muestran a *Alta Vista* como el mejor. Con el tamaño del índice no existe una tendencia clara, ya que se proponen hasta cuatro motores distintos. La presencia de *Google* en los estudios (en los más recientes o actualizados) suele implicar su designación como el de índice más grande, en cambio, cuando *Google* no aparece (seguramente por la fecha del estudio, anterior a su puesta en marcha), *Alta Vista* ocupaba esa posición de privilegio. El número de motores analizados por los diversos estudios es otro parámetro de cierto interés. En la Tabla 4.13 se observa cómo abordan los autores sus análisis:

	Número de motores analizados						**Número de Variables utilizadas**
	<=3	**4 - 6**	**7 - 9**	**10 – 15**	**15 – 20**	**> 20**	
Chu	X						9
Leighton, 1995		X					5
Leighton, 1999		X					1
Arents				X			4
Bharat		X					2
Ralph			X				5
Wishard					X		1
Lawrence		X					4
Gordon			X				5
Gwizdka	X						6
Ming	X						5
Notess						X	5
Ljoslund, 1999	X						1
Ljoslund, 2000					X		5
Thomas, 2002		X					2

Tabla 4.13. Síntesis del número de características empleadas en los estudios experimentales. Fuente: elaboración propia.

[95] Variable estudiada en 15 de los 31 trabajos analizados aunque curiosamente, se trata de la variable de estudio más afectada por la continua evolución de la web.

Una amplia mayoría de estos estudios analiza, como máximo seis motores de búsqueda (situación que se produce en 9 de 14 ocasiones). En cuanto al número de parámetros medidos en cada estudio, el de Chu y Rosenthal (curiosamente el primero) emplea hasta nueve, centrándose la mayor parte de los estudios realizados (un total de 7 sobre los 12 restantes) en analizar cinco o seis variables. Siguiendo la línea marcada por el estudio anterior, donde se analizan varios parámetros en pocos motores, lo normal sería que este grupo mayoritario de estudios analizasen a un conjunto reducido de motores, aunque no existe una tendencia clara en este punto, ya que su cobertura es muy diversa.

Propuestas globales de evaluación.

Existe también otra serie de trabajos que centrada más en la definición de un marco global para la evaluación de los motores de búsqueda, que en el mero análisis comparatorio de la efectividad de unos motores frente a otros. Aunque el objeto principal de estos trabajos varía de uno a otro y resulta complicado establecer una línea común en ellos, sí coinciden en su pretensión de no ofrecer como resultado qué motor es más preciso o cuál de ellos ha crecido más en el último trimestre, sino en la idea de concebir una propuesta integral de evaluación de estos sistemas.
El primero de los trabajos que analizados, el de Oppenheim, presenta una sugerencia de criterios mínimos necesarios a tener en cuenta en el diseño de una metodología de evaluación, fruto de una exhaustiva síntesis de las medidas empleadas en otros trabajos de evaluación de estos sistemas [OPP, 2000]. A este trabajo le sigue un interesante estudio realizado por Savoy y Picard [SAV, 2001], en el que analizan la efectividad de los distintos modelos sobre los que se basan los SRI en la web. Este estudio no analiza el comportamiento de un motor específico frente a otro, sino que estudia la reacción producida al trasladar un modelo diseñado en su origen para el entorno de los SRI tradicionales al nuevo contexto de la web. El tercero de los trabajos analizados, expone la necesidad de encontrar una metodología ajena a los juicios de *relevancia*, basada en unos parámetros de sensibilidad y utilidad de los documentos. Por último, el cuarto trabajo recogido, presenta una propuesta global de evaluación de los SRI en la web elaborada desde el punto de vista del usuario final [JOH, 2001].

Oppenheim (2000).

Los autores de este trabajo, Oppenheim, Morris y McKnight llevan a cabo un estudio exhaustivo de las metodologías aplicadas en la evaluación de los distintos motores de búsqueda por varios autores anteriores[96] y, a partir de la síntesis de estos procedimientos, exponen una serie de parámetros que les permiten formular una metodología de evaluación de los motores de búsqueda. Los métodos más empleados para la evaluación de los motores de búsqueda pueden agruparse en cuatro categorías:

a. Evaluaciones a pequeña escala
b. Evaluaciones basadas en los tests Cranfield
c. Evaluaciones basadas en los tests Cranfield con estimación del tamaño del motor
d. Evaluaciones que eluden la *exhaustividad*

El primer método lo componen estudios que han llevado a cabo un conjunto de *evaluaciones a pequeña escala*. En estos casos, los experimentadores realizan un escaso número preguntas sobre un grupo reducido de motores de búsqueda. A continuación, analizan todos los resultados (aunque su número llegue a ser de varios miles de documentos) y determinan la *precisión*, el solapamiento entre los motores de búsqueda y la *exhaustividad* relativa de cada motor.

En estos ensayos, parten de la hipótesis de que el conjunto de documentos recuperados representa al web en su conjunto. Para los autores, "estos estudios permiten realizar pocas búsquedas y los resultados no tienen por qué ser necesariamente extrapolados" [OPP, 2000]. El número de preguntas realizadas debe ser necesariamente pequeño, porque en caso contrario, los experimentadores deberían revisar miles y miles de documentos y el estudio se alargaría mucho tiempo, y tardarían demasiado tiempo en ofrecer los resultados[97]. El segundo método lo componen las *evaluaciones basadas en los tests Cranfield*. Es fácil deducir que, en la práctica, la mayoría de las evaluaciones realizadas calculan la *precisión* y la *exhaustividad* de cada motor, o alguna medida derivada de las anteriores, tal como es el caso frecuente de la *exhaustividad* relativa. Generalmente, estas medidas se

[96] Gran parte de estos autores han sido recogidos previamente en los apartados de Estudios Explícitos y Estudios Experimentales.

[97] Con lo que, previsiblemente, el mismo estaría condenado a ser considerado obsoleto incluso antes de finalizarse.

calculan a partir de una muestra de documentos recuperados, muestra cuyo tamaño es variable según el estudio, aunque, por regla general, su valor oscila entre los veinte y treinta primeros documentos recuperados por cada motor. El número de motores analizados es muy variable (algunos estudios se conforman con tres motores analizados y otros llegan a la veintena o a la cuarentena) y el número de preguntas tampoco sigue un parámetro estándar. Un tercer método, variante del anterior, lo conforman las *evaluaciones basadas en los tests Cranfield con estimación del tamaño del índice del motor*.

Estos estudios suelen ser de bastante extensión ya que realizan cientos de preguntas a distintos motores de búsqueda y manejan colecciones de resultados de gran magnitud. Para calcular los tamaños de los índices se siguen diversas técnicas. Una de ellas, la más rudimentaria, es identificar el motor que más resultados ofrece sobre una pregunta, contar cuántas páginas relevantes devuelve y luego calcular hasta qué porcentaje de este valor alcanza el resto, extrapolando que (en la realidad de la web), esos porcentajes de tamaño relativo van a mantenerse.

Esta técnica tan simple adolece de varios problemas, porque ¿cómo se puede estar seguro de que el motor que más devuelve puede representar toda la web? y ¿qué pasa con el resto de documentos relevantes que un motor no comparte en común con el motor de mayor volumen?

Otra serie de estudios desarrollan sus hipótesis apoyados en métodos de estimación estadística más consistentes que las anteriores. El último método consiste en *evaluaciones que eluden la exhaustividad*, es decir que evitan hacer uso de esta medida. Varios estudios la obvian con el objeto de no basar sus juicios en una medida algo imprecisa, sino solamente ser inferida a partir de unas estimaciones. No por ello estos estudios dejan de ser completos, ya que la mayoría analiza muchas variables de los motores de búsqueda, tales como: *precisión*, características de la interface, herramientas para la formulación de preguntas, presentación de los resultados, sintaxis empleada y tipos de agentes inteligentes usados en la recopilación de información. Ocasionalmente incluyen la longitud esperada de búsqueda como parámetro a considerar como alternativa a las tradicionales *precisión* y *exhaustividad*. También hay estudios que analizan el contenido de los documentos, entendiendo a este contenido como el tamaño de los documentos recopilados y los tipos de fuentes indexadas[98].

[98] Las fuentes de las páginas web son: URL, sitios telnet, sitios gopher, servidores ftp, títulos de páginas web, sumarios, textos completos y páginas recomendadas por usuarios.

Criterio	**Conclusiones generales**
Número de páginas cubiertas y cobertura	Ningún motor de búsqueda por sí solo indexa toda la web. Los resultados además proyectan algunas dudas sobre la validez de usar la ***exhaustividad* relativa** como medida.
Actualización del índice y número de enlaces erróneos	La frecuencia de actualización es casi tan importante como estimar el tamaño del índice del motor de búsqueda porque proporciona una idea de su adecuado funcionamiento.
Relevancia	Normalmente se hace uso de una visión binaria de la *relevancia* (sí – no), aunque otros introducen escalas.
Sintaxis	Se identifican tres categorías: frases, lenguaje natural y booleanas. Otros incorporan más elementos a esta clasificación, aunque no aportan conclusiones significativas.
Materias	Este campo no está muy estudiado. Es normal que un motor ofrezca mejores resultados en un área que en otra.
La dinámica naturaleza de la web	Una página puede cambiar días después de ser indexada. Esto implica que los resultados ofrecidos por distintos motores sean diferentes. Incluso puede haber desaparecido. Los experimentos han de considera este dinamismo y las evaluaciones deben hacerse en fechas próximas y repetirse a para incorporar los cambios que se vayan produciendo.
Tiempo de respuesta	Parámetro no siempre fácil de calcular, como podría parecer, ya que está muy supeditado al tráfico de la red.
Características diferentes del sistema	Las particulares características de cada sistema influyen considerablemente en sus usuarios. En este campo destacan varios trabajos que recomiendan hacer uso de motores "ad hoc" para cada necesidad particular de información.
Opciones de búsqueda	Se estudia si los motores hacen uso de búsquedas simples y/o avanzadas. Algunos motores emplean otros tipos de búsquedas más complejas. También se estudian las preferencias de búsqueda de los usuarios de los motores.
Factores humanos y cuestiones de la interface	Los motores deben evaluarse bajo una perspectiva de usuario que evite los problemas derivados de la subjetividad. El análisis de la interface lleva implícito una obsolescencia, ante la rápida evolución de estos sistemas. Estos ingenios resultan inaccesibles, en gran medida, para invidentes.
Calidad de los resúmenes	Aspecto que parecía relegado a un papel residual dentro de la evaluación de los motores. Se han comparado resúmenes procedentes de motores y directorios con resúmenes de bases de datos en línea de amplia difusión, comparando la longitud del resumen y la legibilidad. El resultado indica que los directorios poseen mejores resúmenes.

Tabla 4.14. Resumen de características empleadas en las evaluaciones de los SRI. Fuente: Oppenheim.

La Tabla 4.14 recoge el amplio conjunto de criterios empleados en trabajos previos de evaluación y sus conclusiones. Tras realizar esta exhaustiva síntesis de criterios para la evaluación de los motores de búsqueda, Oppenheim, Morris y McKnight consideran que "resulta claro que muchas de las investigaciones desarrolladas son inconsistentes en método y enfoque. Esto demuestra la inmadurez de este campo de estudio" [OPP, 2000]. Para los autores las evaluaciones de los motores de búsqueda deberían incluir, como mínimo, los siguientes criterios:

1. *Precisión*
2. *Exhaustividad* relativa (haciendo uso del Método de Clarke y Willet[99], para su cálculo).
3. Velocidad de respuesta, analizada varias veces al día y calculada en términos de promedio.
4. Consistencia de resultados a lo largo de un determinado período de tiempo.
5. Proporción de enlaces fallidos.
6. Proporción de duplicados.
7. Calidad promedio de resultados
8. Evaluación de la amigabilidad de la interface
9. Calidad de la ayuda
10. Opciones para la visualización de los resultados
11. Presencia de avisos en la pantalla
12. Cobertura (haciendo uso del Método de Clarke y Willet, para su cálculo).
13. Longitud esperada de búsqueda.
14. Longitud y legibilidad del resumen.
15. Efectividad del motor de búsqueda (usando el Método de Back y Summers[100] para su cálculo).

---

[99] Clarke, S.J. and Willet, P. Estimating the recall performance of Web search engines. Aslib Proceedings, 49 (7), 1997, 184-189.

[100] Back, J. And Summers, R. Unpublished results.

Asimismo, los autores consideran que los ensayos deben realizarse haciendo uso de tres tipos de búsqueda: de referencia simple, frases en lenguaje natural y expresiones booleanas.

Savoy y Picard (2001).

Este estudio aborda la evaluación de los SRI en la web desde una perspectiva diferente, en tanto que no analiza a unos motores de búsquedas específicos, sino que evalúa la efectividad de los modelos de recuperación de información en los que se encuentran basados los algoritmos sobre los que se construyen estos SRI, "investigando si las técnicas usadas en los SRI mejoran la efectividad en la recuperación de información cuando se aplican a una colección de documentos web[101]" [SAV, 2001].

Este estudio pone en duda la aplicación de técnicas tradicionales tan comunes como los procedimientos de *aislamiento de la base de las palabras*[102], la presencia de una lista de palabras vacías o la tradicionalmente aceptada importancia de la aparición de las palabras en el título de la página.

Aunque este estudio analiza la realización de los distintos modelos y no ofrece como resultado la designación de un motor de búsqueda como el mejor de los evaluados (tal como hacen los estudios anteriores), sí que resulta interesante conocer qué parámetros emplea para comparar la efectividad de los distintos modelos y qué resultados ofrecen. Además, presenta una serie de datos muy interesantes procedentes de las conferencias TREC sobre el cálculo de la *precisión* en términos de promedio tras analizar los primeros mil documentos devueltos.

En primer lugar, Savoy y Picard miden la *precisión* tras analizar los 10 primeros documentos devueltos por cada modelo y los 20 primeros documentos. Los resultados de este primer análisis son los siguientes:

a. El modelo probabilístico devuelve unos resultados ligeramente superiores a los obtenidos por el modelo del vector clásico (44.9%

---

[101] Esta colección de páginas web está mantenida por TREC y se conoce como WT2g., tiene un tamaño de 2194 MB y consta de 247.491 páginas distintas extrraidas de 969 URLs, con más de un millón ochocientos cincuenta mil términos de indización. Asimismo, contiene 1.171.795 enlaces con una media de 4.73 enlaces por página. Para más información sobre la misma, los autores recomiendan la consulta del trabajo: 'Overview of TREC-8 web track', de Hawking, D. et al. recogido en la publicación 'Proceedings of TREC'8' de 1999.

[102] En inglés "stemming".

frente al 43% con diez documentos analizados y 39.5% frente a 38.65% con veinte documentos analizados).

b. El modelo vectorial que emplea el esquema tf-idf (considera la frecuencia inversa de los términos en los documentos), no arroja resultados muy satisfactorios (entre el 27% y el 23%).

c. Resultan especialmente pobres los resultados del esquema basado en la frecuencia de aparición de los términos en los documentos y el simple coordinate match (ambos modelos ofrecen resultados alrededor del 13% tanto con diez como veinte documentos analizados).

d. El dato más revelador, es el bajo nivel de la *precisión* medida desde el punto de vista de promedio tras analizar una serie de mil documentos, ya que el valor del modelo probabilístico es del 26.68% frente al 23.38% del modelo del vector.

El estudio analiza también los resultados de la *precisión* (medida en términos relativos), cuando se hace uso de los términos recogidos en los títulos de las páginas web y en las zonas de descripción de los documentos[103]. En este caso, mantiene su hegemonía del modelo probabilístico (Okapi), que aumenta ligeramente su distancia sobre el modelo del vector (31.83% frente al 28.26%), aunque descienden considerablemente los valores de *precisión* de todos los modelos, llegándose a situaciones extremas tales como la del modelo basado en la frecuencia de aparición de términos en los documentos, con un descenso del cincuenta por ciento aproximadamente.

Por lo tanto, para Savoy y Picard, en contra de una de las recomendaciones más clásicas que se llevan a cabo a la hora de editar una página web, el insertar una breve descripción de la página en el título, no mejora significativamente la *precisión* en la recuperación de información.

No todos los motores de búsqueda desarrollados hacen uso de procedimientos de aislamiento de la base o de filtrado del conjunto de palabras a almacenar en los índices por medio de una lista de palabras vacías. Este estudio demuestra que el uso de estas técnicas clásicas mejora apreciablemente la *precisión*, ya que, por ejemplo, tanto en el modelo probabilístico como en el del vector, la presencia de las palabras vacías aumenta sus niveles alrededor de un 14% y el procedimiento de aislamiento de la base la mejora un 21% aproximadamente.

---

[103] Los autores no explican qué zonas son las de descripción de un documento. Convencionalmente, esas zonas son conocidas con "meta-etiquetas" de una página web.

Otras conclusiones de este estudio indican que el añadir palabras en las preguntas puede mejorar significativamente la *precisión* promedio y que asignar una mayor importancia (o peso) a las palabras que aparecen en el título o en los encabezados de las páginas, no posee efectos significativos sobre la *precisión*.

Casi de forma paralela a la realización de este estudio, estos mismos autores plantearon un análisis de la efectividad mostrada por tres modelos de recuperación de información en la web: el clásico (representado por el TRECeval software y que siguen la mayoría de los motores de búsqueda del mercado), el de extensión de enlaces (el usado por *Google*) y uno basado en el modelo probabilístico [PIC, 2001].

En este nuevo estudio, aunque los autores siguen sin comparar motores de búsqueda en particular, sí están llevando a cabo una evaluación de las prestaciones de *Google* en toda regla, ya que, a partir de las conclusiones del informe anterior (el modelo probabilístico ofrece mejores resultados de *precisión* que el modelo clásico del vector, representado en este nuevo estudio por el software TRECeval), resulta muy interesante ubicar la posición de *Google*, y por extensión de todos los motores que consideran los enlaces como base del alineamiento de los resultados, en esta clasificación.

Los valores medios obtenidos, cuando se han considerado los enlaces que salen de las páginas analizadas, se recogen en la Tabla 4.15:

TRECeval	**Extensión de Enlaces**	**Probabilístico**	**Probabilístico vs TRECeval**	**Probabilístico vs Ext. Enlaces**
0.253	0.266	0.267	+5.53%	+0.37%

Tabla 4.15 Resumen de resultados de la efectividad de los distintos modelos de recuperación de información en la web. Fuente: Picard, J, and Savoy, J. Searching and Classyfing the web using hyperlinks: A logical approach. [En línea] Neuchatel: III, Universitè, 2001. <http://www.unine.ch/info/Gi/Papers/ictai.pdf> [Consulta: 6 de diciembre de 2001]

Las conclusiones resultan bastante clarificadoras, los resultados que ofrece el modelo probabilístico son, al menos, igual de buenos que los ofrecidos por un SRI en la web cuyo algoritmo de alineamiento hace un uso extensivo de los enlaces y ambos mejoran ligeramente (alrededor de un cinco por ciento), los resultados de *precisión* media frente a los sistemas basados en el modelo clásico del vector.

Schlichting y Nilsen (1997).

Para los autores de este trabajo, las primeras evaluaciones que se realizaron de los motores de búsqueda "poseían un escasísimo nivel científico[104]" [SCH, 1997]. Schlichting y Nilsen indican que los trabajos de Winship y de Leighton de 1995 constituyen un avance importantísimo en este campo de trabajo.

En estos trabajos, los autores se percataron del problema de la presencia de documentos duplicados en las respuestas de los motores de búsqueda, de hecho, Leighton suprimió los documentos duplicados en su análisis.

No obstante, tal como opinan Schlichting y Nilsen, "con la puesta en funcionamiento de motores de búsqueda de gran tamaño (como *Alta Vista*), cuyos índices contienen más de treinta millones de páginas, medir el número de aciertos que proporcionan a una pregunta, no resulta a la larga una medida efectiva, la calidad de los resultados es mucho más importante que la cantidad de documentos que se entreguen como resultado" [SCH, 1997].

El objetivo de este trabajo es mostrar una metodología que pueda emplearse para comparar motores de búsqueda y otro tipo de posibles futuros sistemas inteligentes de recuperación de información.

Esta metodología debe encontrar alguna manera de medir la calidad de los resultados ofrecidos por un motor de búsqueda, los autores proponen hacer uso de la metodología SDA[105], que proporciona dos medidas: *d'* que mide la sensibilidad del motor de búsqueda en hallar información útil y *β* que mide cómo de conservador (o de liberal) es el comportamiento del motor de búsqueda a la hora de determinar qué páginas deben formar parte de la respuesta (mide el grado de flexibilidad del motor a la hora de considerar relevante un nuevo documento analizado).

Este método "supera el alcance del simple cómputo de los aciertos, que sólo emplea una pequeña cantidad de los datos e ignora el amplio contexto de la búsqueda de información.

La metodología SDA proporciona una visión más detallada del rendimiento de los motores de búsqueda, incorporando más información dentro de una visión más global" [SCH, 1997].

---

104 Los autores proponen como ejemplo el trabajo de Arents de 1996, que, tal como hemos visto anteriormente, es un trabajo que ha ido actualizándose desde entonces y se ha transformado de un estudio meramente descriptivo de las características externas de los motores de búsqueda en un trabajo mucho más elaborado.

105 SDA son las siglas de "Signal Detection Analisis".

En la realización del experimento, participaron cinco evaluadores quienes realizaron una serie de cinco búsquedas a los motores examinados. Cada expresión de búsqueda contenía entre cuatro y seis términos y se les había pedido que las diseñaran ellos mismos y que las materias objeto de las búsquedas fueran diferentes a las que ellos solían realizar. Las palabras que formaban parte de las expresiones eran insertadas en las mismas según el orden de importancia que le asignaba el evaluador.

Los motores analizados fueron *Alta Vista*, *Lycos*, *Infoseek* y *Excite*. Se obtuvieron un total de 200 documentos, de los cuales sólo 54 fueron considerados relevantes, siendo *Lycos* el de mayor número.

El primer paso de esta metodología es asignar una categoría a cada uno de los resultados devueltos. Estas categorías se fundamentan en la asignación de *relevancia* de cada uno de estos sujetos, y son cuatro:

1. Acierto: resultado relevante localizado en el experimento
2. Falsa alarma: resultado irrelevante localizado en el experimento
3. Perdido: documento relevante no localizado en el experimento
4. Rechazado correctamente: resultado irrelevante debidamente no localizado en el experimento

El siguiente paso es determinar las ratios de *Acierto* y de *Falsa Alarma* de cada motor de búsqueda. La primera es el resultado de dividir aciertos localizados por cada motor entre el total de aciertos localizados por todos los motores. De forma análoga, la segunda es el resultado de dividir los resultados irrelevantes entregados por un motor entre los enlaces erróneos proporcionados por los cuatro motores.

En la Tabla 4.16 aparecen los resultados obtenidas en este experimento:

	Lycos	***Excite***	**Infoseek**	***Alta Vista***
Acierto	37.3	27.5	23.5	17.6
Falsa Alarma	21.8	25.4	26.8	28.9

Tabla 4.16 Ratios de acierto y falsa alarma en cuatro motores de búsqueda. Fuente: Signal Detection Analysis of WWW Search Engines. [En línea] Seattle: Microsoft, 1997. <http://www.microsoft.com/usability/webconf/schlichting/schlichting.htm> [Consulta: 2 de diciembre de 2001]

Una conclusión rápida de estos resultados indica que *Lycos* presenta mejores resultados, en el conjunto de ambas ratios calculadas, que los otros motores analizados.

Aunque la Metodología SDA no se detiene en este punto, ya que estos valores le sirven para la determinación de la medida de la *sensibilidad* (*d'*) y del la *flexibilidad (β)*.

En el caso de la primera medida, su valores suelen oscilar entre 0 y 2, cuando el valor de la medida *d'* es nulo, indica que el motor es incapaz de discriminar entre documentos relevantes o no relevantes. En la segunda medida, el motor que ofrece el valor más alto es el que se considera más conservador.

Los resultados obtenidos en el estudio asignaron a *Lycos* (con un discreto valor de 0.48 sobre 2), como el motor más sensible, muy por encima de los demás, *Excite* obtiene un valor de 0.07 muy próximo a cero e incuso Infoseek y *Alta Vista* presentan valores negativos. En el caso de la flexibilidad, no se obtienen resultados para aquellos que discriminan negativamente, siendo *Excite* el que obtiene un mayor valor de la medida *β*, lo que implica que es menos flexible que *Lycos*.

El más claro resultado de este estudio es que el rendimiento de los motores de búsqueda está muy lejos de ser considerado ideal. Ninguno de los cuatro motores analizados, se han aproximado a unos niveles aceptables de sensibilidad y de flexibilidad. Estos valores aceptables resultan necesarios para una correcta aplicación de esta metodología, es por ello que los autores comentan: "será necesario que mejore la tecnología para continuar con este tipo de estudios o bien los mismos deben variar su alcance" [SCH, 2001].

Obviando los pobres resultados de este estudio en particular, la gran contribución de este trabajo es la sugerencia de un método objetivo para la evaluación de la efectividad de la recuperación de información.

Johnson, Griffiths y Hartley (2001).

En la *Universidad Metropolitana de Manchester*[106], se ha desarrollado recientemente un estudio destinado a establecer un marco global para la

[106] El estudio se llama 'DEVISE: A framework for the Evaluation of Internet Search Engines', y ha sido desarrollado por el CERLIM ('Centre for Research in Library and Information Management') de la Universidad Metropolitana de Manchester.

evaluación de los motores de búsqueda en Internet. La estructura de este informe se divide en cuatro capítulos.

El primero de ellos proporciona una introducción general a los motores de búsqueda en Internet y a su evaluación. El segundo capítulo muestra de forma gráfica la evolución de estos ingenios en la búsqueda de la mejora de su ejecución. El tercer capítulo revisa las metodologías empleadas en la evaluación de estos sistemas, partiendo desde una perspectiva sistémica hacia un punto de vista centrado en el usuario.

En último lugar, en el cuarto capítulo de este informe se construye un marco global para la evaluación de los SRI en la web, basado en criterios de satisfacción del usuario. En la revisión de las metodologías empleadas hasta el momento para la evaluación de los motores de búsqueda, los autores identifican tres grupos:

a. Evaluaciones basadas en los tests Cranfield.
b. Evaluaciones de sistemas interactivos centradas en el usuario.
c. Evaluaciones basadas en la satisfacción del usuario.

Como comentaban Oppenheim, Morris y McKnight [OPP; 2000], "la mayoría de los estudios que evalúan la realización de los motores de búsqueda están basados en la noción de *relevancia* y por tanto, están concebidos siguiendo un diseño tipo test de Cranfield" [JOH, 2001].

Así, el método de las *evaluaciones basadas en los tests Cranfield* es el más empleado a lo largo de las evaluaciones realizadas. Las medidas más significativas empleadas dentro de este método son la *exhaustividad* y la *precisión*, a pesar de sus problemas de determinación exacta.

Resultan especialmente interesantes las recomendaciones de procurar evitar, en lo máximo, favorecer a un motor frente al resto de los analizados cuando se lleva a cabo un proceso de evaluación. También recomiendan contemplar la naturaleza dinámica de la web y llevar a cabo los ensayos en períodos breves de tiempo para que los resultados puedan ser comparables. Por último, hacen referencia a las posibles dimensiones en las que se puede medir la *relevancia*.

El segundo grupo de métodos es el que los autores denominan *evaluaciones de sistemas interactivos centradas en el usuario*. Johnson, Griffiths y Hartley comentan que muchos autores reniegan del trascendental protagonismo

que el método anterior le confiere a los juicios de *relevancia*, considerando a esta trascendencia como la culpable de los bajos niveles de efectividad que proporcionar normalmente estos estudios.

Este grupo lo conforman aproximaciones alternativas a la evaluación de los SRI que eluden hacer uso de los juicios de *relevancia*. Una de estas medidas alternativas propuestas es la *utilidad* de los documentos devueltos, aunque los autores de este informe recuerdan que la *utilidad* [LAN, 1973] está muy vinculada a la *relevancia*, y más que ayudar a superar los métodos empleados en los *tests de Cranfield* parece que se vuelve a ellos.

No obstante, recientemente sí se ha conseguido determinar una medida simple de la *utilidad* de los documentos basada en cálculos empíricos[107]. La tendencia más seguida, dentro de esta primera medida, es la de presentar los documentos de la respuesta alineados según un criterio determinado de clasificación, de manera que los primeros documentos sean más útiles al usuario que ha formulado la ecuación de búsqueda.

La segunda medida considerada dentro de este grupo es la usabilidad, una medida que "pretende involucrar al usuario aún más en la evaluación" [JOH, 2001].

Es una medida de la capacidad del sistema para operar de forma efectiva, eficaz y satisfactoria. Se ha estudiado desde muchos puntos de vista: exactitud, tasa de error, número de comandos empleados, número de descriptores utilizados, pantallas necesarias para la recuperación de información y percepciones del usuario.

Posteriores investigaciones han analizado las relaciones de estas medias con las características de los usuarios (como por ejemplo, con sus capacidades cognitivas). El contexto en el que se encuentra la persona que va a realizar las búsquedas también debe ser objeto de estudio.

El tercer grupo de métodos empleados para la evaluación de los motores de búsquedas, es el de las *evaluaciones basadas en la satisfacción del usuario*. Una experiencia piloto en este campo agrupó quince medidas en cinco categorías: *relevancia*, eficiencia, utilidad, satisfacción del usuario y conectividad de la página. Un total de once evaluadores participaron en la experiencia realizando diversas cuestiones sobre cuatro motores de búsqueda, cada uno de ellos llevó a cabo sus juicios de *relevancia* (con base en las categorías anteriores).

---

107 SU, L. 'Evaluation measures for interactive information retrieval'. Information Processing and Management, 28 (4), 1992. p: 503-516

Posteriormente se clasificaron los 20 documentos más valorados para cada pregunta y cada evaluador eligió los cinco documentos que para él resultaban más relevantes. Los resultados de esta experiencia piloto muestran tantas diferencias entre los cuatro motores analizados que resulta imposible identificar a uno por encima de los demás.

Otros estudios analizan una serie de parámetros muy vinculados con el usuario final, que se muestran de forma sintetizada en la Tabla 4.17:

Características tangibles	Cómo de bien organizada está la información; los métodos de búsqueda disponibles y la cantidad de información disponible.
Fiabilidad	Consistencia de la sintaxis en las operaciones de búsqueda y *relevancia* de los resultados obtenidos.
Sensibilidad	Entrega rápida de los resultados.
Confianza	Información no duplicada; enlaces válidos y actualizados.
Empatía	Primera impresión; lenguaje natural; existencia de ventanas de ayuda; ejemplos de búsqueda y selección del idioma de los documentos.

Tabla 4.17 Síntesis de parámetros orientados al usuario final. Fuente: elaboración propia.

Los autores afirman que "informes preliminares de la calidad del servicio mostrado por lo motores de búsqueda muestran un nivel bajo" [JOH, 2001]. La causa principal, a su juicio, es la dificultad de refinar una búsqueda de una forma verdaderamente sensible para mejorar la calidad de los documentos entregados al usuario.

La propuesta de un marco global de evaluación considera los criterios del usuario, las medidas a emplear y el contexto donde se desarrollará. Esta propuesta se esquematizaría en Tabla 4.18:

Evaluación de la satisfacción del usuario de los SRI en la web (Johnson, Griffiths y Hartley, 2001)			
Eficacia	**Eficiencia**	**Utilidad**	**Interacción**
P1: *precisión*	Tiempo empleado en la sesión de búsqueda	Valor global de los resultados	Satisfacción del usuario con el formato de salida de los resultados
P2: satisfacción del usuario con la *precisión*	Tiempo de respuesta del sistema	Satisfacción con los resultados	Satisfacción del usuario con las posibilidades de manejo de los resultados
P3: comparación de P1 y P2	Tiempo de establecimiento de la *relevancia*	Grado de resolución del problema	Satisfacción del usuario con la representación de los documentos
R1: el alineamiento de los documentos que hace el sistema		Valor medio de la participación	I1: Satisfacción del usuario con la interface del sistema
R2: satisfacción del usuario con el alineamiento R1		Calidad de las fuentes	I2: Satisfacción del usuario con el formato de la pregunta
R3: comparación de R1 y R2		Números de enlaces seguidos	I3: Satisfacción del usuario con las posibilidades de modificar las pregunta
			I4: Satisfacción del usuario con la claridad global del sistema al preguntar

Tabla 4.18 Propuesta de un marco global de evaluación de motores de búsqueda del Informe Denise. Fuente: Informe Denise.

Dentro de la primera dimensión de la evaluación, la efectividad del sistema, las medidas P2 y R2 se obtienen a partir de una serie de entrevistas que se realizan a los usuarios del sistema, una vez que se han realizado las búsquedas y se conocen los resultados devueltos. Los usuarios valoran tanto la *precisión* como el alineamiento de los documentos devueltos por el motor

de búsqueda y con base en estas valoraciones se determinan los valores de estas dos medidas.

La eficiencia del sistema se mide en términos de tiempo, aunque en este caso se propone diferenciar el tiempo de respuesta del sistema (el parámetro más utilizado de este grupo en otros estudios anteriores), del tiempo que necesita el usuario para el experimento y del tiempo que precisa el usuario para discernir si el documento entregado como resultado le resulta válido o no.

Considerar estos tres tiempos por separado es un aspecto muy interesante de este estudio, en tanto que, implícitamente, se expresa de forma empírica una serie de aspectos que hasta ahora se medían de modo muy subjetivo, tales como el formato de presentación de los documentos y la dificultad de manejo del sistema (si se tarda mucho en conocer si el documento es relevante o no seguramente se deberá a que el SRI es de complicada gestión), y no se relega a esta dimensión a una mera cuestión de cómputo de los segundos y décimas de segundo que un motor de búsqueda necesita para entregarnos sus resultados.

La *utilidad* de los documentos constituye la tercera dimensión que se analiza en esta propuesta centrada en el usuario. El valor medio de la participación de los usuarios ha sido calculado a partir de sus propias opiniones, ya que indicaron (en una escala de 1 a 5) el valor de su participación, con respecto a la información que han obtenido. Los usuarios manifestaron también el grado en el que la información recuperada contribuyó a satisfacer sus necesidades de información. Igualmente, opinaron sobre su satisfacción con los resultados y el valor de los documentos recuperados.

La última dimensión analiza la interacción con el sistema, "este concepto es a menudo citado pero pocas veces ha sido definido. En el contexto de los motores de búsqueda hace referencia a cómo el usuario directamente interacciona con el sistema para recuperar la información que necesita" [JOH, 2001]. En la línea de considerar las características intrínsecas de estos sistemas, llama la atención el establecimiento de cuatro medidas que analizan la interactividad del usuario con la interface del sistema.

Normalmente, en la mayoría de los estudios todos estos aspectos se agruparían en la medida **I1** (satisfacción del usuario con la interface del sistema), pero Johnson, Griffiths y Hartley destacan la importancia de esta interactividad en el momento de la realización de una pregunta (de la expresión de la necesidad de información en términos de una ecuación de búsqueda), para ello definen **I2** (satisfacción del usuario con el formato de la pregunta), **I3** (satisfacción del usuario con las posibilidades que el sistema le

ofrece para modificar la pregunta sin tener que repetir toda la consulta) y por último, la medida **I4** (satisfacción del usuario con la claridad de la vista global del sistema en el momento de realización de una pregunta). También se establecen diferencias entre el formato de visualización de los documentos y el formato de manejo de los documentos, conceptos que tradicionalmente siempre han sido considerados como uno sólo. Este conjunto de medidas fue sometido a una evaluación para estudiar su comportamiento y posible validación. Veintitrés estudiantes del *Departamento de Información y Comunicación* realizaron búsquedas, cuya temática eligieron ellos mismos en tres motores de búsqueda[108], y realizaron tantas modificaciones de la expresión original de la pregunta como ellos consideraron necesarias para obtener el conjunto de documentos deseados. Posteriormente realizaron sobre los resultados una serie de juicios de *relevancia* en términos de la escala: Relevante – Parcialmente relevante – No relevante.

La primera hipótesis a validar en este experimento es que la evaluación de un SRI en la web por medio de medidas centradas en la satisfacción de los usuarios, es un proceso multidimensional con un cierto grado de relación entre las dimensiones analizadas. A partir de los resultados del experimento se puede construir la Tabla 4.19 de correlación entre las cuatro dimensiones de la propuesta:

	Coeficientes de correlación de las dimensiones analizadas[109]			
	Global	Sistema **A**	Sistema **B**	Sistema **C**
Efectividad	**0.759**	**0.779**	**0.795**	**0.729**
Eficiencia	**0.817**	**0.843**	**0.908**	**0.741**
Utilidad	**0.710**	0.362	**0.930**	**0.806**
Interacción	*0.592*	*0.511*	*0.660*	*0.580*

Tabla 4.19 Resultados de la propuesta de un marco global de evaluación de motores de búsqueda del Informe Denise. Fuente: Informe Denise.

[108] Como el objeto de este estudio es validar la propuesta de marco global de evaluación, los autores no indican qué motores fueron empleados en el experimento sino sólo los resultados obtenidos.

[109] Se resaltan en negrita los valores de correlación superiores a 0.7, que los autores entiende como "correlación fuerte". En cursiva se destacan los valores de "correlación moderada".

La eficiencia del sistema es la dimensión que presenta un mayor nivel de correlación, siendo la interacción la que presenta el menor nivel. Este experimento valida la hipótesis anterior, y además "el resultado podría permitirnos sugerir que la eficiencia es el criterio más presente en la mente de los evaluadores, mientras que los usuarios de los sistemas que han participado en este estudio han demostrado escaso interés en la interacción con el mismo" [JOH, 2001].

Dentro de cada dimensión analizaron también los niveles de correlación entre las distintas variables propuestas. Por regla general, los resultados de estas correlaciones fueron de un nivel más bajo que los recogidos en la tabla anterior.

Es decir, es más fuerte el grado de vinculación de las distintas dimensiones empleadas para la evaluación por parte del usuario que la relación que las distintas variables mantienen en el seno de cada grupo analizado. Los resultados más destacados de este segundo proceso de análisis presentan que la satisfacción con la *relevancia* de los documentos posee un nivel mucho más alto de correlación que la satisfacción con el alineamiento y que la satisfacción con los resultados obtenidos y el valor conferido a estos resultados en su globalidad posee mayor correlación que las otras variables relacionadas con la utilidad.

Con lo cual, este estudio, viene a representar de forma empírica, las principales necesidades de un usuario de un SRI, el usuario necesita información de valor que cubra sus necesidades de información y la misma le debe ser entregada en un espacio de tiempo pequeño.

Tablas e ilustraciones.

Tabla 4.1 Clasificación de los motores de búsqueda según su tamaño. Fuente: Notess, G. R. Comparing Internet search engines [En línea]. Charles Sturt University, 1999. <http://www.csu.edu.au/special/online99/proceedings99/103a.htm> [Consulta: 10 noviembre 2001] 113

Tabla 4.2 Clasificación actualizada de los motores de búsqueda según tamaño y determinación de diferencias entre tamaños estimados y tamaños declarados. Fuente: Search Engine Statistics: Database Total Size Estimates. [En línea]. Bozeman, MT: Notes.com, 2000. <http://www.searchengineshowdown.com/stats/sizeest.shtml/> [Consulta: 12 marzo 2002]114

Ilustración 4.1 Clasificación de los motores de búsqueda según tamaño. Fuente: Search Engine Sizes. [En línea]. Darien, CT: INT Media Group, 2001 <http://www.searchenginewatch.com/reports/sizes.html> [Consulta: 10 de febrero de 2002]115

Ilustración 4.2 Evolución en el tiempo del tamaño de los motores de búsqueda. Fuente: Search Engine Database Change Over Time. [En línea]. Bozeman, MT: Notes.com, 2000 <http://www.searchengineshowdown.com/stats/change.shtml> [Consulta: 10 de marzo de 2002] 116

Ilustración 4.3 Evolución en el tiempo del tamaño de los motores *Alta Vista* y *Google*. Sullivan, D. Media Metrix Search Engine Ratings [En línea]. Darien, CT: INT Media Group, 2000 [Consulta: 10 de febrero de 2002] 118

Tabla 4.4 Doce de los veinticinco sitios más visitados de Internet. Fuente: NetRatings/Nielsen....119

Tabla 4.5 Millones de búsquedas por día en distintos motores. Fuente: Searches Per Day. [En línea]. Darien, CT: INT Media Group, 2001 <http://www.searchenginewatch.com/reports/perday.html> [Consulta: 12 de marzo de 2002] 120

Ilustración 4.4. Accesos de múltiples sitios web. Fuente: StatMarket Search Engine Ratings. [En línea]. Darien, CT: INT Media Group, 2000. < http://www.searchenginewatch.com/reports/statmarket.html/> [Consulta: 24 de noviembre 2001] 121

Tabla 4.6. Sinopsis de estudios explícitos de evaluación de los motores de búsqueda. Fuentes: elaboración propia. 123

Tabla 4.7. Porcentajes de fallos en las referencias ofrecidas por los motores de búsqueda, Fuente: Notess, G. R. Search Engine Statistics [En línea]. Bozeman, MT: Notes.com, 2000. <http://www.searchengineshowdown.com/stats/> [Consulta: 10 noviembre 2001] 138

Tabla 4.8. Resultados del estudio de Ljoslund de búsqueda de palabras raras. Fuente: Lojslund, M. Evaluation of twenty Web search engines on ten rare words ranking algorithms. [En línea]. Trondheim: Sør-Trøndelag University, 2000. <http://www.aitel.hist.no/~mildrid/dring/paper/Comp20.doc> [Consulta: 21 de noviembre de 2001] 140

Tabla 4.9. Resultados del estudio de Ljoslund de búsqueda de palabras raras ordenados por la *precisión*. Fuente: Lojslund, M. Evaluation of twenty Web search engines on ten rare words ranking algorithms. [En línea]. Trondheim: Sør-Trøndelag University, 2000. <http://www.aitel.hist.no/~mildrid/dring/paper/Comp20.doc> [Consulta: 21/11/2001.....141

Tabla 4.10. Tabla resumen del estudio de Ljoslund de búsqueda de palabras raras ordenados por la *precisión*. Fuente: Lojslund, M. p.6 [Consulta: 21/11/2001] 141

Tabla 4.11. Tabla resumen del estudio de Sarah Thomas. Fuente: C | NET 143

Tabla 4.12. Sinopsis de estudios experimentales. Fuente: elaboración propia. 144

Tabla 4.13. Síntesis del número de características empleadas en los estudios experimentales. Fuente: elaboración propia. 145

Tabla 4.14. Resumen de características empleadas en las evaluaciones de los SRI. Fuente: Oppenheim. 149

Tabla 4.15 Resumen de resultados de la efectividad de los distintos modelos de recuperación de información en la web. Fuente: Picard, J, and Savoy, J. Searching and Classyfing the web using hyperlinks: A logical approach. [En línea] Neuchatel: III, Universitè, 2001. <http://www.unine.ch/info/Gi/Papers/ictai.pdf> [Consulta: 6 de diciembre de 2001] 153

Tabla 4.16 Ratios de acierto y falsa alarma en cuatro motores de búsqueda. Fuente: Signal Detection Analysis of WWW Search Engines. [En línea] Seattle: Microsoft, 1997. <http://www.microsoft.com/usability/webconf/schlichting/schlichting.htm> [Consulta: 2 de diciembre de 2001].....155

Tabla 4.17 Síntesis de parámetros orientados al usuario final. Fuente: elaboración propia. 159

Tabla 4.18 Propuesta de un marco global de evaluación de motores de búsqueda del Informe Denise. Fuente: Informe Denise. 160

Tabla 4.19 Resultados de la propuesta de un marco global de evaluación de motores de búsqueda del Informe Denise. Fuente: Informe Denise. 162

BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES

5 Propuesta del modelo de evaluación.

RESUMEN: La evaluación de los sistemas de recuperación de información en la web precisa de una propuesta multidimensional de análisis de su funcionamiento. A partir de una perspectiva inductivista, se procede a la enunciación de una propuesta integral de evaluación que contemplará aspectos relacionados con la relevancia de los documentos, cantidad y calidad de los índices de los motores de búsqueda, modalidades de recuperación de la información y similitud entre estos índices. Previamente se establecen los criterios de selección de la muestra de motores objeto de estudio y se define una función discreta de determinación de la relevancia de los documentos en función de su posición en la respuesta. Toda propuesta de evaluación de la web debe prever las posibles contingencias típicas de este entorno, de naturaleza tan dinámica como particular, por lo que deben definirse a priori las acciones a tomar en cada situación identificada.

Propuesta del modelo de evaluación.

Introducción.....167

Planteamiento de un Procedimiento de Evaluación. 171

Selección de la muestra objeto de estudio. 171

Determinación del tipo de búsqueda.....172

Establecimiento de criterios para analizar las páginas web.....172

Verificación de la pertinencia de los documentos. 172

Agrupamiento de los índices.....174

Objeto de la evaluación. 174

Desarrollo del Procedimiento de Evaluación. 175

Fase 1: Selección de la muestra objeto de estudio.....175

Fase 2: Elección del tipo de búsqueda y formulación de la pregunta.....181

Fase 3: Establecimiento de criterios previos al análisis de los resultados.....184

Fase 4: Análisis de los resultados. 191

Fase 5: Normalización de la precisión y de la exhaustividad. 191

Fase 6: Determinación de la similitud. 192

Fase 7: Agrupamiento de los índices. 199

Ejemplo de evaluación. 199

Tablas e ilustraciones. 207

Introducción.

A lo largo de los capítulos anteriores se han presentado, de forma suficientemente amplia, los conceptos generales sobre los SRI y su evaluación (tanto de los tradicionales como de los que funcionan en la web), analizándose pormenorizadamente el conjunto de medidas que se han propuesto para considerar la efectividad de los mismos.

Si bien el conjunto de variables empleados en la evaluación de los SRI en la web es cuantioso y heterogéneo (se han identificado más de treinta en los estudios y posteriormente en la presentación de las metodologías de evaluación ha surgido otra importante cantidad), se ha podido

comprobar[110] que las medidas aplicadas en la mayoría de los estudios, se sintetizan en cuatro grupos: (1) las capacidades de búsqueda y la interface del sistema, (2) la precisión y la exhaustividad (calculadas normalmente en términos de promedio), (3) el tamaño del índice del motor de búsqueda y (4) el grado de solapamiento y número de errores detectados en los motores.

En algunos trabajos interesantes se han presentado una serie de aportaciones para la evaluación de los SRI en la web. Curiosamente, algunas de estas medidas, en contra de lo que cabría suponer por su originalidad e importancia, no han trascendido más allá de los trabajos en los que se incluyen o bien han quedado difuminadas por el conjunto total de aportaciones propuestas.

La primera de estas contribuciones es la propuesta de Leighton y Srivastava de diseñar una función de precisión que considere la posición de los documentos en la respuesta del motor, favoreciendo la presencia de documentos relevantes entre los veinte primeros [LEI, 1997]. Con esta función, los autores no sólo están evaluando la precisión del sistema sino que, al mismo tiempo, evalúan el funcionamiento del algoritmo de alineamiento del sistema, es decir, cómo presenta los documentos recuperados tras una operación de búsqueda.

Ya se ha indicado en el capítulo anterior, que esta medida de Leighton y Srivastava, sin dejar de ser una medida basada en los juicios de relevancia, muestra una serie de características que la aproxima al área de las medidas centradas en el usuario y, a lo largo de la revisión realizada en los capítulos anteriores [KOR, 1997], [JOH, 2001], ha quedado patente la creciente importancia que vienen adquiriendo estas medidas en los últimos años.

El estudio realizado en los laboratorios de investigación de la empresa NEC, dirigido por Lawrence y Giles [LAW, 1998], aporta la interesante sugerencia de calcular el tamaño estimado de la unión de los índices de dos motores (*Hot bot* y *Alta Vista*), comparándolo a continuación con el resultado de las posibles uniones del resto de pares de índices de motores analizados en el estudio.

El tamaño de los índices ha sido objeto de estudio prácticamente en la mitad de los trabajos analizados en el anterior capítulo, pero, en este trabajo

[110] A partir de las sinopsis de los dos grandes grupos de estudios que hemos identificado en el capítulo anterior (explícitos y experimentales), y de la propuesta metodológica de Oppenheim [OPP, 2000]

en particular, lo que se considera de especial importancia es precisamente esta agrupación (unión en este caso) de los índices de los motores.

Esta asociación de índices de los motores de búsqueda nos traslada a otra cuestión suficientemente merecedora de estudio, que no es otra que preguntarse hasta qué punto debería extenderse esa unión de índices con el propósito de confeccionar un nuevo índice, mucho más extenso, que permita, sino localizar toda la documentación recopilada en la web[111], al menos garantizar unos altos niveles de efectividad entre los primeros veinte o treinta documentos recuperados.

Lawrece y Giles, sin pretender realizarlo en su estudio, cuando estiman que el total de documentos no duplicados producto de la unión de los índices de los dos motores era el total de documentos de la web, introducen un nuevo problema, cuya solución permitiría diseñar una serie de sugerencias a los usuarios de la web del tipo: "si desea localizar documentos sobre una materia ALFA, debe consultar los veinte primeros documentos que ofrecen el motor A, el motor B y el motor C, y no es necesario consultar más documentos en cada motor o probar con otro porque estos tres motores ofrecen unos valores muy altos de precisión y de exhaustividad que difícilmente van a mejorar por mucho que se consulte en otro sitio".

Evidentemente, esta serie de enunciados se construiría a partir de los resultados de una experimentación previa, por lo que constituiría un interesantísimo conjunto de recomendaciones debidamente fundamentadas, muy cercanas al usuario y alejadas, en un considerable grado, de problemas de subjetividad (relacionados muchas veces por la costumbre de uso de un motor o por el diseño de la interface de usuario).

Una de las últimas autoras presentadas, Ljoslund expone dos elementos muy importantes en su trabajo [LOJ, 1999]: la *relevancia parcial* y el uso de una *función de similitud*. Cuando Ljoslund comenta la necesidad de hacer uso de varias medidas para valorar el grado de relevancia de un documento, está presentando una de las situaciones más normales que se le presenta a un usuario cuando tiene que valorar la relevancia de un documento.

Si el usuario sólo dispone de la posibilidad de asignar un valor de "relevante" a aquel documento que considera pertinente a una pregunta y un valor de "no relevante" para aquellos que no poseen esta característica, va a verse

---

[111] Teniendo en cuenta que los índices de los motores difícilmente llegan a recopilar una mínima parte de la web en los casos más favorables, y el grado de solapamiento de su contenido, sería muy complicado encontrar esa unión que recopilara toda la web.

obligado a llevar a cabo una serie de juicios con los que verdaderamente no va a estar de acuerdo, en tanto que localizará documentos totalmente relevantes y otros que son relevantes pero en menor medida y, al no disponer de una escala de valores más amplia, van a resultar finalmente igual de relevantes.

Esta situación, muy frecuente, es una de las causas más empleada por toda la serie de autores que manifiestan su disconformidad por el uso de los juicios de relevancia en los procesos de evaluación de los SRI. Una solución interesante sería ampliar la escala de asignación de valores de relevancia de un documento, por ejemplo, pasar de la escala binaria a una escala de cinco puntos (de 0 a 4, siendo 0 el valor que le correspondería a un documento no relevante y 4 al documento más relevante). Otra posibilidad, más simple, es la que propone Ljoslund, insertar un valor de 0,5 para la *relevancia parcial*.

Donde Ljoslund resulta verdaderamente innovadora es en la propuesta de emplear una *función de similitud* (en su caso emplea la función de Jaccard) [LOJ, 2000], para determinar "en qué grado las respuestas de un motor forman un subconjunto de las respuestas de otro". Con este cálculo, Ljoslund verdaderamente persigue establecer un nuevo método de determinación del grado de solapamiento existente entre los índices de los motores de búsqueda, tomando como base las respuestas que estos motores llevan a cabo a las mismas preguntas, además de examinar la semejanza entre el conjunto de documentos gestionados por estos sistemas y el orden en que los mismos le son presentados al usuario.

Finalmente, hay que tener presente los resultados del trabajo de Johnson, Griffiths y Hartley, quienes concluyen que "la evaluación de un SRI en la web basada en características centradas en la satisfacción de los usuarios, es un proceso multidimensional donde existe un cierto grado de relación entre las dimensiones analizadas" [JOH, 2001].

Si bien las medidas centradas en el usuario resultan algo difíciles de determinar y de trasladar al contexto de la web, estas conclusiones promueven la idea de que la evaluación debe realizarse desde un punto de vista multidimensional que permita calcular, por medio de un determinado conjunto de procedimientos, una serie de medidas que representen la efectividad del sistema de recuperación de información. Esta serie de medidas, deben contemplar los siguientes aspectos:

- El *contexto* de la web: su tamaño y su capacidad de cambio. Estos dos aspectos obligan a que los procedimientos que se realicen no sean excesivamente complejos en tanto que el estudio precisa de

sucesivas actualizaciones para reflejar verdaderamente la efectividad de los sistemas.

- Las *prestaciones* de los sistemas: no todos los sistemas poseen las mismas posibilidades de recuperación de información y, además, no parece recomendable permitir que la cobertura y el tamaño de los índices de los motores influya considerablemente en su evaluación, por lo que el conjunto de motores evaluados debe ser más o menos homogéneo.
- La *utilidad* de los documentos para los usuarios: no es suficiente que un documento sea relevante, además ha de ser útil, es decir *pertinente* [LAN, 1973].
- El *alineamiento* de los resultados: un sistema va a resultar, a la larga, más satisfactorio para sus usuarios cuanto mejor sea el alineamiento que lleva a cabo de los resultados de una búsqueda, porque los usuarios agradecen encontrar los documentos relevantes entre los primeros devueltos.
- Una *visión multidimensional* de la evaluación, que permita establecer el resultado de un proceso de análisis a partir de la interacción de varias ratios.

Planteamiento de un Procedimiento de Evaluación.

Con el objeto de satisfacer, en la medida de lo posible, los aspectos comentados en el apartado anterior, debemos plantear un procedimiento de evaluación de la efectividad de los SRI en la web que lleve a cabo el siguiente conjunto de tareas.

Selección de la muestra objeto de estudio.

Con el fin de obtener una adecuada visión, han de evaluarse de forma conjunta motores de búsqueda que posean unas características similares, tanto en el tamaño de sus índices como en las posibilidades de búsqueda. Del mismo modo, se deben tener en cuenta las restricciones geográficas que, en algunos casos se presentan en los motores, siendo más conveniente confrontar conjuntos de motores sin ningún tipo de restricción. La selección del idioma también es importante, en tanto que "algunos motores de búsqueda prácticamente no pueden ser susceptibles de evaluación en un idioma distinto del inglés". [GWI, 1999]

Determinación del tipo de búsqueda.

Debe determinarse qué tipo de búsqueda se va a evaluar en los motores. Normalmente, estos sistemas proporcionan dos opciones: *simple* y *avanzada*. La primera de ellas permite recuperar documentos donde aparecen algunos de los términos que forman parte de la expresión de búsqueda del usuario, aunque hay sistemas (como *Google*), que emplean por defecto la intersección de los subconjuntos de documentos asociados a cada término, con lo que sus resultados deben poseer los tres términos solicitados. Esta opción de intersección forma parte de las búsquedas avanzadas en la mayoría de los motores. Otra posibilidad de búsqueda, corriente en la mayoría de los motores, es la combinación de expresiones booleanas con la localización de frases exactas, modalidad de búsqueda que, en principio, debe proporcionar mejores niveles de efectividad, aunque esta hipótesis debe ser verificada. Por lo tanto, en este punto se debe clarificar el grupo de sistemas a evaluar y el tipo de búsqueda (o los tipos) que se van a emplear en esta evaluación. Las preguntas que se empleen en los procesos de evaluación han de estar redactadas de antemano y, en la medida de lo posible, deben ser formuladas de forma clara y precisa.

Establecimiento de criterios para analizar las páginas web.

Antes de proceder a la evaluación de la relevancia de los documentos integrantes de la muestra a analizar, deben quedar fijados los procedimientos a seguir cuando se presenten situaciones características de la web, tales como la posibilidad de encontrarnos páginas de contenido similar alojadas en distintos sitios, o que la respuesta a una pregunta contenga muchas referencias al mismo sitio web (situación muy frecuente en el caso de los portales dinámicos) o el documento al que se enlaza nos remite a otra dirección distinta de la que posee el motor en su base de datos. Otra circunstancia que debe clarificarse de antemano es la escala de determinación de la relevancia: binaria (relevante o no relevante), relevancia parcial [LOJ, 1999], etc.

Verificación de la pertinencia de los documentos.

A partir de una muestra suficientemente significativa del resultado de una búsqueda, pueden estimarse los valores de *precisión* y *exhaustividad relativas* [SAL, 1983], necesarios para el establecimiento de una medida de valor simple [FRA, 1997] [VOI, 1997] [RIJ, 1999]. El tamaño de la muestra también va a tener relación con el grado de similitud de los vectores de

resultados, ya que, en principio habrá más coincidencias conforme mayor sea el tamaño de la muestra elegida.

Por lo tanto, la elección de este tamaño no debe llevarse a cabo sin haber verificado su incidencia en la determinación de estos valores y considerando el hecho adicional de que, si se optara por una muestra relativamente pequeña (alrededor de los diez primeros documentos devueltos por cada motor), los pares de valores E-P serían bastantes más altos[112] que los que se devuelven en estudios que han aumentado el tamaño de la muestra [SAV, 1999], y si el tamaño de la muestra es relativamente grande, entonces surge el problema de la dificultad de repetir el experimento con asiduidad, en tanto que se precisa de más tiempo para su realización.

Análisis de la similitud de los resultados.

La lista de direcciones que ofrece un motor puede entenderse, dentro de una perspectiva de espacio vectorial, como un vector de direcciones que puede compararse con otra lista de direcciones (o vector) producida por otro motor en respuesta a la misma pregunta.

Esta comparación se puede llevar a cabo por medio de una función de similitud, como por ejemplo la *función del coseno* [SAL, 1983], [BAE, 1999], con valores comprendidos entre 0 y 1 según la similitud de los vectores de resultados.

Existen otras funciones de similitud, igualmente válidas para la representación de esta semejanza entre los resultados [SAL, 1983], [LOJ, 2000]. La determinación de esta analogía nos proporcionará información sobre la composición de los índices de los distintos motores y sobre su función de alineamiento, ya que si dos motores muestran unos altos grados de similitud (una vez comparados los resultados de varias búsquedas), quiere decir que en sus índices existen muchos documentos comunes y que, además, sus algoritmos de alineamiento han respondido de forma similar.

Este cálculo presenta dos modalidades, según se consideren o no los documentos que no son relevantes con la necesidad de información del usuario. Si se consideran, el análisis permite extraer conclusiones más amplias sobre el contenido de los índices de los motores, ya que igual que habrá

[112] Existe un alto grado de posibilidades de que algún motor devuelva diez documentos relevantes en las diez primeras respuestas, lo que proporciona unos pares de valores E-P muy superiores a la realidad.

coincidencias entre documentos relevantes parece lógico suponer que habrá también coincidencias entre los documentos que no son relevantes.

El grado de similitud de los vectores de resultado se presenta como una medida más completa que el grado de solapamiento de ambos motores[113]. En este mismo proceso, de forma previa al cálculo de la similitud de los vectores de resultados, se pueden establecer los porcentajes de enlaces erróneos y de enlaces duplicados que aparecen en la respuesta de cada motor, como factor adicional de medida de la calidad de esa respuesta.

Agrupamiento de los índices.

A la finalización del experimento, se verificará si existe una tendencia establecida de agrupamiento hacia un valor central de las similitudes calculadas entre todos los motores de búsqueda. En caso afirmativo, cabrá la posibilidad de aplicar alguna técnica de agrupamiento de estos índices, con el fin de confeccionar un índice más completo que pudiera proporcionar unos mejores valores de efectividad que los calculados para cada motor por separado.

Objeto de la evaluación.

Este proceso de evaluación posee dos objetivos principales: (1) asignar un valor de efectividad a las operaciones de recuperación de información que recoja, en la medida que esto sea posible, las características intrínsecas del entorno donde se desarrollan esos procesos (la web) y, (2) determinar el grado de similitud de los resultados ofrecidos por estos SRI.

Esta efectividad, medida en términos de valor único, no se basará exclusivamente en los juicios de relevancia a los que se someta a los documentos recuperados, sino que integrará otra serie de aspectos típicos del entorno (tamaño del índice de los motores de búsqueda, presencia de resultados duplicados en los vectores de resultado, enlaces erróneos o inactivos, páginas de portales dinámicos producto de consultas a bases de datos, etc.) que han sido tratadas en otros procesos de evaluación de forma aislada, impidiendo la determinación de su verdadera incidencia en la efectividad de la recuperación de información.

[113] Evidentemente, ambos factores se encuentran muy relacionados.

El segundo objetivo de este proceso de evaluación persigue concretar hasta qué punto los distintos sistemas de recuperación de información ofrecen prestaciones similares en el tratamiento de los contenidos informativos que gestionan.

En este caso no se trata de ponderar la efectividad o eficiencia del sistema, o la amigabilidad o potencia de la interface de búsqueda, en este caso se pretende establecer con qué grado de semejanza (o de disparidad) opera cada uno de estos ingenios. Si además de establecer estos grados, se pueden tomar como raíz de un proceso de agrupamiento, se estaría sentando las bases para la construcción de un SRI sensible al contexto, al usuario y dotado de unos altos niveles de efectividad.

Tanto la medida de la efectividad de la recuperación de información (diseñada desde una óptica globalizadora e integradora de las características peculiares de la web) como el establecimiento de la similitud de los contenidos gestionados por cada motor, nos permitirá disponer de una visión multidimensional de la evaluación de estos sistemas, mucho más completa que la mayor parte e estudios analizados en el capítulo anterior.

Desarrollo del Procedimiento de Evaluación.

A continuación se desarrollarán las distintas tareas de las que consta el procedimiento de evaluación, para lo cual se han propuesto las siguientes fases:

1. Selección de la muestra objeto de estudio
2. Elección del tipo de búsqueda y formulación de la pregunta.
3. Establecimiento de criterios previos al análisis de los resultados.
4. Análisis de los resultados.
5. Normalización de la precisión y de la exhaustividad.
6. Determinación de la similitud.
7. Agrupamiento de los índices.

<u>Fase 1: Selección de la muestra objeto de estudio.</u>

A lo largo de esta primera etapa se ha de determinar el idioma a emplear en las expresiones de búsqueda que vamos a utilizar en el experimento, a continuación debemos establecer el conjunto de motores a analizar y por último, considerar si se van a introducir restricciones geográficas al ensayo.

Selección del idioma.

El idioma elegido para la realización de este ensayo es el Español. La elección de nuestra Lengua se debe a dos tipos de razones: la novedad que supone el desarrollo de una propuesta de análisis de la efectividad de la recuperación de información en nuestra Lengua y el menor volumen de documentos gestionados por los motores (en comparación con otros idiomas más extendidos). La casi totalidad de estudios revisados emplean el Inglés como el idioma para la construcción de las expresiones de búsqueda, por lo que, con nuestra elección, aportamos una pequeña variante a todo el conjunto de evaluaciones realizadas con anterioridad.

En segundo lugar, creemos que un deficiente funcionamiento del algoritmo de indexación o del algoritmo de búsqueda y alineamiento de un motor, puede camuflarse gracias a la gran cantidad de documentos que conforman sus índices, lo que le permitiría entregar como respuesta extensas listas de documentos (más o menos pertinentes) a pesar de un bajo nivel de funcionamiento de esas rutinas. Si el conjunto de documentos gestionado es menor, existe la posibilidad de que estas deficiencias influyan más en la calidad de la respuesta, permitiendo al evaluador tomar debida nota de ello. Por otro lado, el idioma a elegir tampoco puede ser un idioma que posea un escaso número de documentos publicados en la web, ya que el evaluador corre el riesgo de no poder localizar para una pregunta el mínimo de documentos que deben conformar la muestra a analizar, tal como le sucedió a Gwizdka y Chignell [GWI, 1999].

En el caso de nuestra Lengua, este problema no parece que vaya a presentarse (aunque habrá que descartar a algún motor generalmente evaluado en trabajos anteriores con escasa cobertura lingüística más allá del Inglés), ya que, según datos de la propia *Real Academia Española*[114], aproximadamente un 5% de las páginas web están escritas en Español, lo que representa un porcentaje suficientemente amplio para no correr el riesgo anteriormente expuesto. De hecho, tal como recoge el *Anuario del Instituto Cervantes* del año 2000: "las estadísticas sobre el uso de la lengua en la red, aunque con manifiestas discrepancias en detalles que podemos considerar secundarios, nos llevan a una situación muy conocida, con el Inglés como lengua general de comunicación, con cerca de un 60 por

[114] Datos aportados por Antonio Cid, en la presentación del **II Congreso Internacional de la Lengua Española**, celebrado del 16 al 19 de octubre de 2001 en Valladolid. Fuente: Noticias.com [En línea] <http://www.noticias.com/noticias/2001/0110/n01101212.htm> [Consulta: 2 de diciembre de 2001]

ciento de uso, en el menos favorable de los censos, mientras que el 40 por ciento restante cada vez está más fraccionado entre otras lenguas. Según datos de julio de 1999, si se incluye el idioma Inglés, el Español se usa en un 4,3 por ciento de las conexiones, el Inglés en un 57,4 por ciento y ninguna otra lengua alcanza un 10 por ciento del uso total. Si excluimos al idioma Inglés, el Español se usa en un 10,1 por ciento de las conexiones a Internet que se realizan en todas las demás lenguas. Chino, Alemán y Japonés serían las únicas que contabilizan magnitudes superiores" [MAR, 2000].

Conjunto de motores a analizar.

Si la selección del idioma constituye un paso fundamental para el establecimiento del procedimiento de análisis, no menos trascendente es la elección del conjunto de motores a analizar. Para su determinación se pueden seguir varios criterios:

1. Revisar trabajos anteriores y elegir aquellos que hayan sido más veces evaluados. Así, los resultados de nuestro estudio podrían compararse fácilmente. El problema de este criterio es que puede dejar fuera a motores de reciente puesta en funcionamiento (WISEnut por ejemplo), ya que su presencia en los estudios revisados es menor (*Alta Vista* o *Lycos*, por ejemplo).
2. Analizar los motores de mayor popularidad entre los usuarios de Internet. El problema de este criterio es el posible sesgo que introduzcan estos estudios, generalmente realizados en los Estados Unidos, donde la mayoría de búsquedas no se realizan en Español.
3. Elegir un conjunto de motores que posean entre sus índices una cantidad de documentos significativa en nuestro idioma y que, al mismo tiempo, presenten características de búsqueda similares. Este criterio permitirá evitar ensayos donde resulte difícil determinar una muestra válida para su posterior análisis.

Si se sigue el tercer criterio, que a priori no registra una deficiencia del nivel de los dos anteriores, hay que dilucidar cómo estimar el número de documentos que indexa cada motor en Español, o por lo menos establecer una clasificación. Esta estimación del porcentaje se ha llevado a cabo por medio de la realización de diez *grupos de búsquedas* sobre un conjunto de nueve[115] motores de búsqueda. Cada *grupo de búsqueda* consiste en la localización de los documentos que un motor contiene sobre una temática

[115] En principio se pretendía trabajar con un número mayor de motores, pero se fueron desechando algunos por diversos motivos, que enumeraremos posteriormente.

de carácter general, sobre la cual se realizaron cuatro interrogaciones (por ejemplo, para estimar la presencia de documentos sobre la temática "ciudades medianas de España", se realizaron cuatro búsquedas por motor, interrogándoseles por documentos que contuviera la palabra: "Alicante", luego la palabra "Granada", posteriormente la palabra "Málaga" y finalmente la palabra "Murcia". A continuación se suman todos los documentos devueltos por cada motor y se calcula el porcentaje que sobre la temática posee cada uno de ellos (en el caso de la anterior temática, el motor *All the Web* devuelve el 15% de los 5735 documentos localizados, mientras que los motores *Excite*, Microsoft Network y *Terra* alcanzan valores alrededor del 7% cada uno de ellos). Tras consultar la documentación de cada motor, se procedió a establecer las ecuaciones de búsqueda en idioma Español sobre las cuestiones recogidas en la tabla 5.1 sin considerar ningún otro tipo de restricción adicional.

Tema	**Cuestiones**
Capitales de provincia de mediano tamaño	Bilbao, Sevilla, Valencia y Zaragoza
Capitales de provincia de pequeño tamaño	Alicante, Granada, Málaga y Murcia
Escritores españoles	Miguel de Cervantes, Lope de Vega, García Lorca y Juan Ramón Jiménez
Personajes famosos de la Historia	Julio César; Cristóbal Colón; Hernán Cortés y Simón Bolívar
Científicos famosos	Isaac Newton, Marie Curie, Albert Einstein y Luc Montagnier
Pintores españoles	Diego de Velásquez, Francisco de Goya, Salvador Dalí y Pablo Picasso
Políticos famosos contemporáneos	Felipe González, Hill Clinton, Tony Blair y Jacques Chirac
Periodistas españoles	Diego Carcedo, Jesús Hermida, Luis Carandell y Julia Otero
Deportistas famosos	Zinedine Zidane, Miguel Indurain, Severiano Ballesteros y Pau Gasol
Áreas especializadas de la Ciencia	Biología Celular, Cirugía Cardiovascular, Química Orgánica y Física Teórica

Tabla 5.1 Cuestiones empleadas para medir la distribución porcentual de los distintos motores con base en una serie de preguntas en idioma Español. Fuente: Elaboración propia.

A la hora de realizar las preguntas se siguieron una serie de normas con los nombres propios, que fueron introducidos con la primera letra en mayúscula respetando las tildes (por ejemplo: "Salvador Dalí"). Otra consideración fue evitar localizar documentos por un término que pudiera encontrarse escritos

en varios idiomas ("Javier " y "Xavier"). Se emplearon nombres propios no demasiado corrientes, con el fin de no introducir documentos en la respuesta que no tienen nada que ver por la persona buscada (así, en la localización de documentos sobre periodistas, se prefirió localizar documentos sobre "Diego Carcedo" en lugar de "José María García", al ser la combinación de nombre y apellidos del segundo periodista más común que la del primero).

En la medida de lo posible, se evitó hacer uso de nombres de personas excesivamente largos y que incluyeran preposiciones, artículos o conjunciones (como por ejemplo: "José Ramón de la Morena" no fue empleado en el ensayo de localización de documentos que mencionaran a periodistas famosos), ya que algunos motores (especialmente *Excite*) tienen problemas con las palabras vacías en los términos de búsqueda. El resultado final de este experimento proporciona la distribución que se observa en la siguiente ilustración:

BIBLIOTECA VIRTUAL

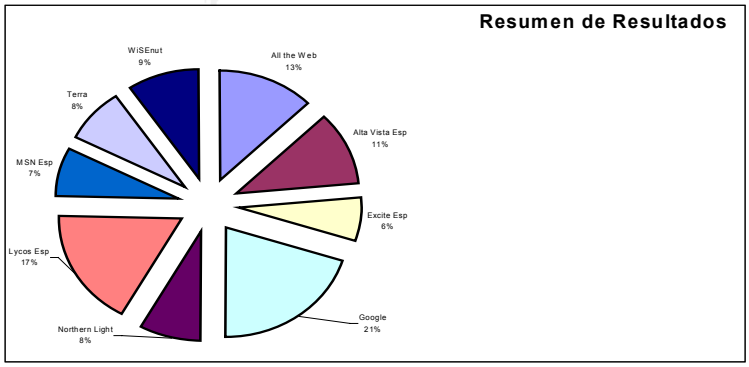


Ilustración 5.1 Distribución porcentual de los distintos motores con base en una serie de preguntas en idioma Español. Fuente: Elaboración propia.

Los motores analizados en este ensayo han sido los siguientes: *All the Web*, *Alta Vista*[116], *Excite*[117], *Google*, *Northern Light*, *Lycos*[118], *Microsot Network*

---

[116] Se evaluaron *Alta Vista* España (<http://es-es.altavista.com>) y *Alta Vista* Global (<http://www.av.com>); eligiéndose el primero porque devolvía más documentos.

[117] Se evaluaron *Excite* España (<http://www.*Excite*.es>) y *Excite* Network (<http://www.*Excite*.com>), verificándose que devuelven la misma cantidad de documentos. De hecho, el primer sitio es una interface de consulta a la base de datos global de este motor para documentos escritos en idioma Español.

[118] Se evaluaron *Lycos España* (<http://www.lycos.es>) y *Lycos* Network (<http://www.*Lycos*.com>); eligiéndose el primero porque devolvía más documentos.

(MSN)[119], *Terra* y *WiSEnut*. En un inicio, también formaron parte del ensayo los motores *Overture* y *Hot bot*, pero fueron descartados porque no informan del total de documentos localizados, precisamente el objeto principal de este experimento. También se consideró inicialmente evaluar el motor *Infoseek*, pero el mismo es actualmente una versión del nuevo motor *Overture* y las búsquedas que se realizan desde su sitio web son devueltas al usuario como referencias de *Overture*[120], por lo que se descartó igualmente. El resultado del ensayo nos permite extraer dos conclusiones importantes que deben tenerse en cuenta en el resto del proceso de evaluación:

- El porcentaje sobre el total devueltos por cada motor de búsqueda analizado, todos superan el cinco por ciento, lo cual nos parece un porcentaje suficientemente representativo para formar parte de la muestra del estudio.
- Los resultados reflejan claramente la diferencia de tamaño existente entre los índices de los distintos motores. Esta circunstancia exhorta a incorporar, en el proceso de evaluación, algún mecanismo corrector de las posibles incidencias del tamaño de un índice con respecto a otro. Este mecanismo de corrección debe introducirse en el cálculo de toda la serie de medidas que se empleen sobre la anterior muestra de motores de búsqueda, con el fin de establecer su eficacia.

Restricciones geográficas.

El objetivo general de nuestro experimento es analizar la efectividad de los motores de búsqueda en Internet, a través de una serie de experimentaciones en las que emplearemos nuestro idioma (el Español), para realizar una serie de búsquedas en los motores indicados en el apartado anterior. En la realización de estas pruebas no resulta conveniente introducir restricciones geográficas, porque pretendemos valorar la efectividad del sistema no valorar los contenidos que se producen en un área geográfica u otra.

---

[119] Al igual que en el caso de *Excite*, se analizaron la versión española (<http://www.msn.es>) y la versión global (<http://www.msn.com>) y se verificó que proporcionaban similar número de documentos.

[120] Como se explica en el siguiente capítulo, el motor *Overture* ha ido absorbiendo a distintos motores, tales como *Infoseek*, *Goto* y posteriormente *Excite*. Este proceso ha coincidido en el tiempo con la realización de nuestros experimentos, lo cual ha motivado que hayamos ido adaptando nuestro trabajo a la evolución del contexto.

Fase 2: Elección del tipo de búsqueda y formulación de la pregunta.

Tipo de búsqueda.

Es sabido que cada motor de búsqueda interpreta de una manera distinta las expresiones que se le introducen desde el modo de búsqueda simple, por lo que tendremos que emplear la modalidad de búsqueda avanzada, asegurándonos así que nuestras comparaciones se realizan entre búsquedas de similar naturaleza. Generalmente, cuando un usuario de un SRI quiere recuperar documentos con algo más de precisión que con la búsqueda simple, suele indicarle al motor qué términos de la expresión de búsqueda desea que aparezcan en el resultado o bien si quiere que algunos términos aparezcan en el resultado tal cual figuran en la expresión (lo que se conoce como "búsqueda por frase literal").

Estas dos posibilidades de mejora no son excluyentes y pueden emplearse al mismo tiempo, proporcionando normalmente mejores resultados que por separado, especialmente en el caso de la "búsqueda por frase literal" que resulta muy exclusiva y de poca exhaustividad. Ahora corresponde dilucidar qué tipo de búsqueda se va a utilizar en el experimento, que pueden ser de cualquiera de los tres tipos siguientes:

- la búsqueda por frase literal
- la búsqueda por intersección
- combinación de ambas
- combinación de ambos tipos de búsqueda y presencia de otros operadores booleanos

Inicialmente se descarta la búsqueda por frase literal (de forma exclusiva) por su poca exhaustividad[121], y se realiza un experimento destinado a analizar la relevancia de los treinta primeros documentos recuperados sobre un mismo tema, empleando las restantes tres modalidades de búsqueda sobre un mismo tema. El tema de la búsqueda es "localización de documentos que informen sobre el alquiler de apartamentos en Málaga (provincia)" y se van a comparar los resultados obtenidos en seis motores de búsqueda; *Alta Vista*, *All the Web*, *Google*, *Lycos*, *Northern Light* y *Terra*.

[121] En Google la búsqueda sobre apartamentos en Málaga devuelve 2790 documentos con la operación de intersección y 6 documentos con la búsqueda por frase literal.

El primer tipo de consulta se basa en la intersección de los subconjuntos con términos de la expresión, el segundo tipo de consulta combina la intersección con la búsqueda de frase literal y el tercer tipo es similar al segundo, pero, en este caso, se empleará (cuando sea posible) el operador de la unión para combinar diversas frases literales de significado similar.

Se han formulado y realizado las operaciones de búsqueda para cada uno de los seis motores de búsqueda y como ejemplo de las expresiones, se pueden observar las empleadas en el motor *Google* (tabla 5.2).

1: Intersección	alquiler apartamentos Málaga
2: Intersección + Frase literal	"alquiler de apartamentos" Málaga
3: Intersección + Frase literal + Sinónimos	("alquiler de apartamentos") ("Costa del Sol" OR Málaga)

Tabla 5.2 Sintaxis empleada en Google en el experimento de análisis de la efectividad de la recuperación de información según el tipo de búsqueda. Fuente: Elaboración propia.

Tras analizar los documentos recuperados por cada tipo de búsqueda, se estimaron la exhaustividad y la precisión en los diez y en los primeros treinta documentos. Este cálculo se pares de valores E-P se ha llevado a cabo por medio del método de Salton de cálculo normalizado de estas variables, obteniéndose los resultados que se observan en las Tablas 5.3:

Altavista	1	2	3	***Google***	1	2	3
30 docs				**30 docs**			
Precisión	0,8981	0,8658	0,8792	Precisión	0,8045	0,901	0,9462
Exhaustividad	0,5546	0,5486	0,5555	Exhaustividad	0,5159	0,528	0,5512
10 docs				**10 docs**			
Precisión	0,8904	0,8904	0,9154	Precisión	0,8875	0,9354	1
Exhaustividad	0,5333	0,5333	0,5444	Exhaustividad	0,575	0,5555	1

Tabla 5.3 (I) Pares de valores E-P obtenidos en el experimento destinado a analizar la efectividad de la recuperación de información según el tipo de búsqueda empleado. Fuente: Elaboración propia.

All the Web	1	2	3	***Lycos***	1	2	3
30 docs				**30 docs**			
Precisión	0,842	0,5186	0,5595	Precisión	0,8362	0,6641	0,7813
Exhaustividad	0,5305	0,4955	0,53125	Exhaustividad	0,5208	0,6285	0,645
10 docs				**10 docs**			
Precisión	0,8904	0,6286	0,5664	Precisión	0,9154	0,9084	0,9452
Exhaustividad	0,6	0,9	0,675	Exhaustividad	0,5444	0,6714	0,625
Northern Light	1	2	3	***Terra***	1	2	3
30 docs				**30 docs**			
Precisión	0,8241	0,7056	0,6315	Precisión	0,752	0,7835	0,628
Exhaustividad	0,5409	0,515	0,5372	Exhaustividad	0,5947	0,5507	0,524
10 docs				**10 docs**			
Precisión	0,6142	0,8464	0,7235	Precisión	0,55	0,6971	0,7235
Exhaustividad	0,5777	0,6142	0,6	Exhaustividad	0,5375	0,55	0,6

Tabla 5.3 (II) Pares de valores E-P obtenidos en el experimento destinado a analizar la efectividad de la recuperación de información según el tipo de búsqueda empleado. Fuente: Elaboración propia.

Los resultados muestran que cuando se consideran los diez primeros documentos, sí parece interesante emplear la tercera modalidad (intersección más frase literal y uso de sinónimos cuando sea posible), aunque cuando se analizan los treinta primeros documentos la situación cambia (sólo *Google* presenta mejores valores), mientras que los demás motores devuelven peores resultados. Así, elegimos el primer tipo de búsqueda, la basada en la intersección para la realización de nuestro estudio y como contraste, optamos por elegir el tercer tipo de búsqueda, cuando la misma sea posible[122].

[122] No todos los motores analizados permiten efectuar búsquedas de la tercera modalidad, y además, en todos los casos no se precisa del operador de unión. Por ello, no existen diferencias muy significativas entre el uso del segundo y tercer tipo de modalidad de búsqueda, porque muchas veces se refunden en una.

Formulación de la pregunta.

Antes de comenzar a realizar las búsquedas, se ha explicado por escrito el objeto de cada pregunta y qué contenidos informativos resultaban válidos para satisfacer la necesidad de información. Esto se ha hecho con el objeto de no ir cambiando el criterio de análisis de la relevancia de los documentos recuperados, evitando introducir dosis de subjetividad en el experimento. En cada búsqueda se ha rellenado un formulario (la siguiente tabla recoge el caso de la búsqueda de apartamentos para alquilar en Málaga), donde se manifiesta qué se espera encontrar y qué puede considerarse útil a efectos de recuperación de información.

Evaluador	FRANCISCO JAVIER MARTÍNEZ MÉNDEZ	**Fecha**	04-11-2001
Pregunta	ALQUILER DE APARTAMENTOS EN MÁLAGA	**Hora inicio**	11:30 am
Motor	*ALTA VISTA*	**Hora finalización**	11:57 am
Ecuación de búsqueda	+alquiler +apartamentos +Málaga		
Enlaces erróneos	2	**Enlaces repetidos**	4
Contenidos válidos	Páginas web que permitan alquilar un apartamento, estudio o vivienda dentro de la provincia de Málaga. Páginas de enlaces o directorios temáticos que permitan encontrar agencias o particulares que alquilen apartamentos.		
Contenidos no válidos	Páginas de inmobiliarias que se dediquen a construir apartamentos en Málaga. Páginas de periódicos que informen de la subida de precio de los alquileres. Páginas con normas municipales sobre arrendamientos.		

Tabla 5.4 Formulario empleado para indicar a priori el objeto de cada búsqueda del experimento. Fuente: Elaboración propia.

Fase 3: Establecimiento de criterios previos al análisis de los resultados.

En esta etapa corresponde decidir cuál es la escala de relevancia que vamos a adoptar a la hora de analizar los documentos y cómo se va a reaccionar (en el curso de este proceso de análisis), cuando se presenten situaciones tales como la repetición de páginas en sitios distintos o se nos devuelva un enlace inactivo.

Todo este tipo de criterios deben establecerse a priori, con el fin de no desvirtuar el proceso de evaluación, ya que, si fuéramos tomando decisiones a medida que se nos presenta el problema a resolver, estaríamos, en cierto modo, proponiendo un método muy influenciado por el contexto y por el motor indexado.

Grados de relevancia.

En el capítulo segundo se ha comentado que adoptábamos como nuestra la idea de Lancaster sobre la relevancia, es decir: "un documento va a resultar relevante cuando su contenido tiene que ver con el objeto de la pregunta y, además, le resulta útil" [LAN, 1973].

Si en el ejemplo de la búsqueda de información sobre el alquiler de apartamentos en Málaga, un motor devuelve una noticia aparecida en un periódico local indicando que los apartamentos turísticos de Málaga y Costa del Sol han subido sus precios un 5% para esta temporada, esa información, relacionada sin duda con el objeto de la pregunta, no resulta útil, ya que no indica en ningún momento cómo y dónde alquilar un apartamento.

En este caso no cabe hablar de relevancia (ni total ni parcial), sino de no relevancia. Es por ello que se adopta la postura de considerar la escala binaria para medir la relevancia, si el documento es relevante (de temática relacionada y útil), la relevancia de este documento valdrá la unidad, mientras que si no es de temática relacionada o bien no resulta útil, la relevancia será nula.

Adicionalmente, se introduce una *función de ponderación de la relevancia,* establecida en virtud de la situación del documento dentro de la respuesta, es decir, de su alineamiento. Para la mayor parte de los usuarios de Internet resulta mejor que un documento se encuentre entre los diez primeros devueltos por un motor como respuesta a una pregunta, que en entre la posición undécima y vigésima, ya que los usuarios van a valorar mejor a un sistema que le proporcione las respuestas correctas antes que las respuestas erróneas.

Este tipo de función puede definirse de forma *continua* como la Leighton y Srivastava [LEI, 1997] o bien de forma *discreta*, es decir, estableciendo diversos intervalos para esta función. Así, si el documento es considerado relevante y se encuentra situado entre los diez primeros documentos del vector resultado, su función de relevancia valdrá la unidad, mientras que si se encuentra situado entre las posiciones undécima y vigésima del vector resultado, su función de relevancia valdrá 0,9 y si se encontrara situado entre

la posición vigésimo primera y trigésima, su función de relevancia valdrá 0,8. En esta misma progresión, el valor de la función de relevancia será 0,7, 0,6 y 0,5 (sucesivamente), si la posición del documento se encuentra situado en la decena de documentos que comienzan a partir de las posiciones 40, 50 y 60 (respectivamente) de cada vector de resultados.

En el caso de que la muestra a analizar fuera aún mayor, establecemos que la función de relevancia de un documento situado en posiciones tan retrasadas en el vector resultado, valdrá 0,5, ya que rara vez un usuario accede a la séptima pantalla de resultados para localizar un documento. En la siguiente tabla se presenta el recorrido de esta función de ponderación de la relevancia.

i: posición del documento en el vector resultado	**$R(d_i)$: Relevancia del documento**
1 <= i <= 10	$R(d_i) = 1$
11 <= i <= 20	$R(d_i) = 0,9$
21 <= i <= 30	$R(d_i) = 0,8$
31 <= i <= 40	$R(d_i) = 0,7$
41 <= i <= 50	$R(d_i) = 0,6$
51 <= i <= 60	$R(d_i) = 0,5$
I > 60	$R(d_i) = 0,5$

Tabla 5.5 Función de ponderación del grado de relevancia de un documento en función de su alineamiento. Fuente: Elaboración propia.

Situaciones características de la web.

Ya se han comentado las diversas situaciones que se presentan a la hora de analizar documentos web, la mayoría propias de este nuevo contexto. Su análisis precisa de la determinación de unos criterios a seguir a la hora de analizarlas, con el objeto de que la evaluación sea lo más cercana a esta realidad. Con vistas a nuestro proceso de evaluación, se han identificado las siguientes incidencias que pueden presentarse en la web y se ha propuesto, previo proceso de reflexión, una decisión a tomar en cada caso:

a. El motor devuelve en una misma respuesta una URL genérica y una URL específica.

Universidad de **Murcia**
... de primer y segundo ciclo, Doctorado, Títulos propios... Alumnos y Titulados Información de interés para estudiantes y titulados de la **Universidad** de **Murcia**. ...
Descripción: Página principal de la **universidad**. Estudios, investigación, servicios, etc. Buscador y directorios.
Categoría: World > Español > ... > Universidades > España > Murcia > Universidad de Murcia
www.um.es/ - 11k - En caché - Páginas similares

Biblioteca de la **Universidad** de **Murcia**
Universidad de **Murcia** ... Desde 1/01/2001, es usted el visitante número : ...
Descripción: Consultas del catálogo, bases de datos, revistas, periódicos y material disponible.
Categoría: World > Español > ... > España > Murcia > Universidad de Murcia > Servicios
www.um.es/biblio/ - 16k - En caché - Páginas similares
[Más resultados de www.um.es]

Ilustración 5.2 Sección del resultado de una búsqueda realizada en el motor Google. Se muestran dos recursos del mismo sitio web, uno específico y otro más general. Fuente: Elaboración propia.

En esta situación, se elige la URL más relacionada con la búsqueda efectuada. Si en el ejemplo de la Ilustración 5.2 se desea recuperar documentos relacionados con la "Biblioteca de la Universidad de Murcia", la URL resultado será **<http://www.um.es/biblio/>** y ésta será la que se someta al juicio de relevancia. Otro caso parecido es cuando, por ejemplo, deseamos información sobre una librería y el motor nos devuelve como dirección genérica la página principal del sitio en Internet de esa librería y como dirección específica una página de ese mismo sitio en la que nos presentan las novedades del catálogo que ya están a la venta. Siguiendo similar criterio, se elige la URL más relacionada con la pregunta, en este caso la primera.

b. El motor devuelve una URL dinámica producto de una consulta a una base de datos.

Se elige la dirección del sitio principal, es decir, la del dominio donde se aloja el portal. Si por ejemplo, la URL localizada es **<http://www.goldenmile.es/esp/prop.asp?ref=GM-4>**, se selecciona **<http://www.goldenmile.es/>** para llevar a cabo el juicio de relevancia.

c. El motor devuelve un enlace inactivo (los típicos errores tipo 400 del protocolo http).

Se considera que el motor ha devuelto un recurso no relevante. Así se penaliza al motor por no haber actualizado su índice suficientemente,

incorporando esta circunstancia en las medidas de precisión y exhaustividad que se obtendrán en una fase posterior del estudio. Así, el valor de la medida de valor simple obtenida a partir de estas dos ratios, incluirá también el porcentaje de enlaces inactivos presentes en la muestra analizada, completando los juicios tradicionales de relevancia.

d. El motor devuelve una URL que en realidad es una redirección a otro sitio web.

El hecho de que un motor posea en sus índices una URL que redirecciona a un nuevo sitio web, puede deberse a dos razones: la indexó cuando la página se encontraba activa (y con todo su contenido) o la ha recopilado a partir de sus "metaetiquetas" que, además de los parámetros de redireccionamiento pueden poseer información sobre su contenido (por ejemplo, la metaetiqueta "http-content"). En el primer caso, es un problema de actualización de índice que no provoca error al acceder al documento, ya que remite a otro documento activo (no cabe hablar de enlace inactivo propiamente). En el segundo caso, el motor ha realizado correctamente su trabajo, ya que ha extraído toda la información posible de esa página. Ahora bien, ¿qué URL debe evaluarse? La respuesta ofrece pocas dudas, la URL redireccionada, ya que realizar un juicio de relevancia en una página que no contiene información (sino "metainformación") resulta ilógico, dejando aparte que el evaluador no dispondría de suficientes elementos de juicio. Por tanto, se considera que la URL válida es la redireccionada, ya que es verdaderamente donde está la información.

e. El motor devuelve un recurso por duplicado.

En este caso se pueden tomar dos alternativas. La primera es considerar la relevancia del recurso la primera vez que aparezca y olvidarnos del resto de sus apariciones, lo que implicaría analizar tantos documentos de más como duplicados se detecten. La segunda alternativa es considerar como fallo grave esta situación y penalizar al motor que la cometa considerando no relevantes a todas las URL repetidas, así sólo se analizarán los documentos objeto de la muestra elegida y el motor que presente muchos duplicados resultará penalizado en la evaluación. Siguiendo la idea de integrar en una medida única de la efectividad

todas las incidencias que se detecten en la evaluación, especialmente las típicas de la web, optamos por esta segunda alternativa, que favorecerá a aquellos motores donde la incidencia de este problema sea menor. Por lo tanto, la medida de valor simple que se obtenga a partir de los valores de precisión y exhaustividad recogerá, además de la ratio de enlaces inactivos, en cierta medida, la ratio de enlaces duplicados.

f. El motor devuelve distintas secciones de un mismo directorio.

En este caso se procede igual que en el caso de los portales dinámicos, se somete a juicio de relevancia la URL del sitio principal y, cuando aparezca una segunda incidencia se sigue el mismo procedimiento que hemos indicado en el ítem anterior, es decir, la segunda aparición de un recurso (la repetición) se considera como URL no relevante.

g. El motor devuelve varias páginas estáticas pero del mismo portal.

En este caso se entiende que cada página es un recurso independiente y que el motor ha hecho bien recopilando estos sitios, por tanto son válidas para nuestro estudio ambas URL.

h. El motor devuelve la misma página en sitios distintos, es decir, mismo contenido en distintas URL.

Aquí se presentan dos alternativas claramente diferenciadas y suficientemente justificables ambas. La primera es considerar que esos dos sitios son el mismo recurso y por tanto cabe considerar al segundo recurso como duplicado (y por tanto, no tenerlo en cuenta). La segunda opción sería considerar que el motor ha realizado dos operaciones de recopilación de información correctas y que no debe entrar a valorar los motivos que llevan a una empresa a disponer de varios sitios alternativos en la red, en este segundo caso se considerarían ambas direcciones como válidas. Como el objeto último de esta evaluación es medir la capacidad del sistema de recuperación, consideramos que esta segunda opción es la más adecuada, ya que el sistema ha recopilado recursos relacionados con un tema, y no ha de entrar en la autoridad o pertenencia de las páginas a una organización. Además, el evaluador

difícilmente podrá decidir cuál es el recurso original y cuál es el recurso duplicado, y además, en el proceso de cálculo de la similitud podrían aparecer ciertas incongruencias si se consideraran duplicados.

i. El motor en lugar de devolver una página sobre un tema, devuelve una página con muchos enlaces a otras páginas con contenidos sobre un tema (las famosas listas de enlaces o de sitios recomendados).

Si esta página verdaderamente da acceso, como mínimo, a un documento relevante, se considera como relevante, ya que nos sirve para localizar la información deseada.

j. El motor devuelve una página índice.

Resulta frecuente encontrar la página índice de un sitio web, recopilada con denominaciones diferentes. Por ejemplo, es fácil que un motor tenga recopilada la URL **<http://www.um.es/index.htm>** y otro tenga recopilada la URL **<http://www.um.es/>**, haciendo referencia ambas direcciones al mismo recurso. Las razones pueden ser varias, especialmente si la página ha pasado a formar parte de la base de datos del motor por medio de la alimentación manual (ya que dos usuarios distintos han podido insertar dos URL diferentes), o bien si la página forma parte del motor gracias a que figura como enlace recomendado en otra página web (el autor de la página ha podido escribir la URL en un formato diferente al empleado en el otro motor). Independientemente de las causas que originan esta situación, resulta evidente que debemos adoptar una normalización de la nomenclatura de las páginas índice de sitios web, para poder comparar correctamente los vectores resultados. También es posible encontrar páginas índice de sitios web sin el tradicional "www" en la denominación del dominio (**<http://libertysurf.es/>** y **<http://www.libertysurf.es/>**). En este caso no cabe hablar de dos URL distintas sino de un caso de redireccionamiento, por lo que se emplea una de las dos para nuestro estudio. Otra situación que puede presentarse, relacionada con las páginas índice de un sitio web, es cuando el motor devuelve el marco principal de esa página índice. En este caso, menos frecuente que los anteriores, pero posible, se actúa con el mismo criterio, es decir, se elige la URL de la página índice.

k. Enlaces preferentes o patrocinados.

Algunos motores proporcionan enlaces preferentes o patrocinados, es decir, que su presencia en la respuesta no se debe a la operatoria del mismo, sino a la inclusión manual de los mismos por parte de los administradores, a cambio de alguna contraprestación. Estos enlaces no van a ser tenidos en cuenta en el proceso de evaluación.

Fase 4: Análisis de los resultados.

Tras la realización de la ecuación de búsqueda y selección de las URL a evaluar, hay que valorar la relevancia de la página recuperada. En el proceso de selección de las URL se debe tomar nota de cuántos enlaces erróneos y duplicados devuelve el motor para una consulta. Una vez la lista de resultados se encuentre debidamente preparada (es decir, sin enlaces erróneos a analizar y sin documentos duplicados), es cuando se debe iniciar el juicio de la relevancia, siguiendo para ello las consideraciones previas recogidas en el formulario. Si un documento es relevante, se le asignará el valor de la función de ponderación de relevancia correspondiente. La información de esta evaluación, pasará a formar parte de una base de datos, donde se almacena la siguiente información para cada pregunta y motor: número de enlaces erróneos, número de enlaces duplicados, relación completa de URL recuperadas y valor de la función de relevancia asignada a cada URL. A partir de estos datos, se pueden calcular las siguientes medidas para cada pregunta y motor: precisión, exhaustividad[123], porcentajes de enlaces erróneos y de enlaces duplicados. En esta fase se calculará también la medida de valor simple de Borko [FRA, 1997], [RIJ, 1999], [VOI 1999], para disponer de un valor simple que permita cotejar la efectividad de una búsqueda realizada en un motor con la realizada en otro.

Fase 5: Normalización de la precisión y de la exhaustividad.

La medida I_1 de Borko obtenida en la fase anterior, si bien incorpora matices intrínsecos del contexto en el que nos desenvolvemos (la web), aún no representa fidedignamente la efectividad de los SRI, ya que, tal como se ha comentado anteriormente, el experimento está influenciado por las

[123] Los pares de valores E-P se calculan según el método de Salton para valores relativos.

diferencias de tamaño de los índices de los motores. Por lo tanto, *resulta conveniente y necesario tipificar los valores obtenidos en la fase anterior con respecto al volumen de documentos contenidos en cada motor*. Se ha estimado anteriormente una distribución porcentual de cada uno, que muestra claras diferencias de tamaño (más o menos constantes), lo que conlleva a suponer que las diferencias de tamaño constituyen un hecho irrefutable y que, de alguna manera, incide en la calidad de los sistemas de recuperación. Como en cada operación de búsqueda podemos extraer una distribución porcentual de las apariciones de los términos relacionados con la expresión en los distintos motores, los valores de esta distribución servirán de base para la extracción de los módulos necesarios y podremos obtener unas medidas de precisión y exhaustividad tipificadas conforme a esa distribución. Como tanto el tamaño del índice de un motor y la precisión (o la exhaustividad) son dimensiones de naturaleza distinta y por lo tanto, no susceptibles de comparación, debemos proceder a la tipificación de los valores que ofrecen cada una de estas dos variables para cada motor, obteniendo así una medida tipificada, que vamos a denominar *precisión normalizada*. Siguiendo un razonamiento similar determinaremos también una *exhaustividad normalizada*. Con estos valores normalizados, se calcularán las medidas de valor simple I_1 e I_2, para verificar si se ha producido alguna corrección sobre el resultado obtenido con las variables sin tipificar.

Fase 6: Determinación de la similitud.

El conjunto de medidas anteriores proporciona una visión muy completa de la efectividad de la recuperación de información. También permite valorar la calidad del índice de documentos, centrándose especialmente en el análisis de su frecuencia de actualización y en la presencia de duplicados. Una vez tomadas estas medidas en los motores objeto de evaluación, es cuando podemos calcular la similitud existente entre los distintos vectores de resultado propuestos por cada motor, estableciéndose también el valor de la similitud media de todo el conjunto de motores analizados.

Función de similitud del Coseno.

Para determinar la similitud existente entre dos vectores, elegimos la función de similitud del Coseno propuesta por Salton y Harman [SAL, 1983], [HAR, 1992]. En el modelo se hace la suposición básica de que la distancia relativa entre dos vectores en el espacio n-dimensional, representa la diferencia entre los perfiles que se han utilizado para configurar dichos vectores. La

función de similitud, función coseno viene dada por la expresión del ángulo formado por sus dos vectores asociados:

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

Ilustración 5.6 Fórmula de la función de similitud del Coseno. Fuente: 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

La dimensión del espacio vectorial es igual al número total de términos utilizado en la caracterización del problema (en este caso, el número total de respuestas del cuestionario). Donde P es el vector problema de un caso almacenado en la Base de Casos y N es el vector problema del nuevo caso. El resultado de la función coseno proporciona valores entre 0 y 1. En el caso de que el valor de Cos (P,N) fuese 1, indicaría que ambos casos (P,N) son idénticos y por lo tanto que los dos motores vectores de resultados son similares. Si se considera el siguiente ejemplo de dos vectores P y N, cuyas componentes se presentan a continuación: vector P = {1; 1; 1; 1; 0; 0; 0; 0} y vector N = {0; 0; 1;1; 0; 0; 1;1}. La Tabla 5.6 refleja el cálculo de la similitud:

P	**N**	**Escalar**	**P²**	**V²**
1	0	0	1	0
1	0	0	1	0
1	1	1	1	1
1	1	1	1	1
0	0	0	0	0
0	0	0	0	0
0	1	0	0	1
0	1	0	0	1
	Sumas	2	4	4
	Coseno	0,5		

Tabla 5.6 Cálculo de la función de similitud del Coseno para el ejemplo de los vectores P y N. Fuente: Elaboración propia.

En la Tabla 5.6, la tercera columna es el producto escalar de los vectores P y N, la cuarta y quinta columna recoge el resultado de elevar al cuadrado los componentes de los vectores P y N (que al tratarse de ceros y unos produce similar resultado que los vectores originales, aunque esta situación no siempre será así). En la fila de "Sumas" se totalizan los sumatorios del producto escalar y de los cuadrados de los vectores y se calcula la función del coseno (según la fórmula de la Ilustración 5.6), obteniéndose un valor de 0.5. Baeza-Yates [BAE, 1999] ilustra gráficamente el significado de esta función de cálculo de la similitud de dos vectores, tal como se muestra en la Ilustración 5.7:

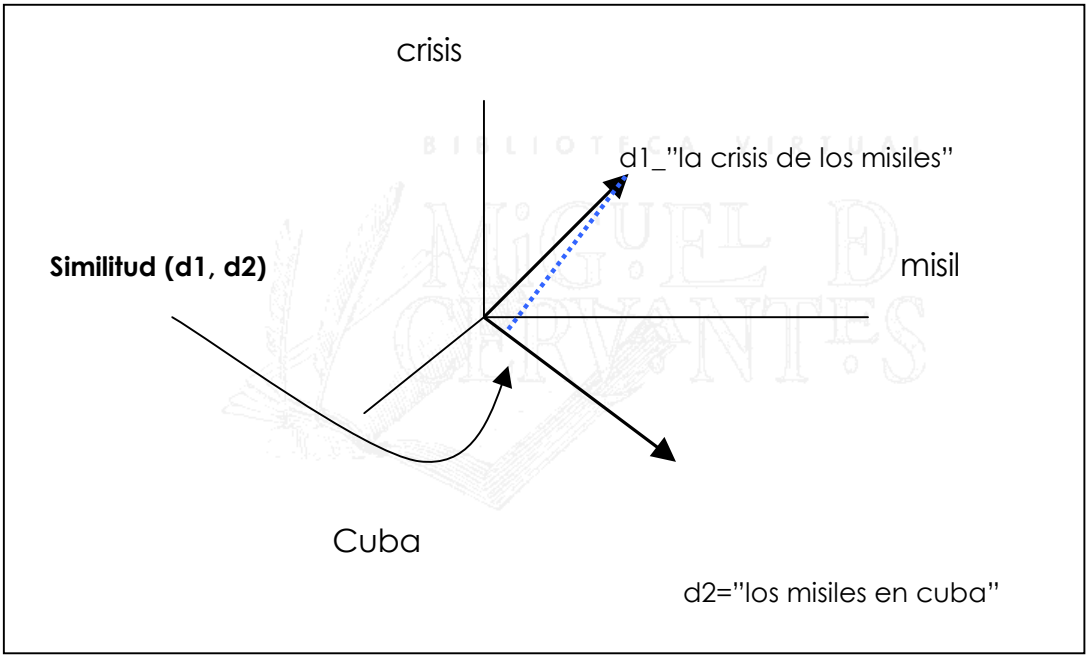


Ilustración 5.7 Significado de la función de similitud del coseno. Fuente: Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999 XX, 513 p. ISBN 0-201-39829-X

<u>Procedimiento de cálculo de la similitud de dos vectores en nuestro experimento.</u>

Con el fin de poder calcular correctamente la función de similitud del coseno en nuestro experimento, se ha de realizar una pequeña adaptación del espacio dimensional en el que trabajamos. Supongamos que se ha realizado una búsqueda con el motor *Google*, a la que corresponde el vector resultado que presentamos en la segunda columna de la Tabla 5.7 y que se va a denominar VR1 ("vector de resultados 1"). Cada URL englobada en este vector resultado va a ser sometida a juicio de relevancia y se le

asignará un valor con base en la función de ponderación de la relevancia que se ha definido anteriormente. El conjunto de valores de relevancia del vector resultado obtenido por los documentos que devuelve *Google* es la columna tercera de la Tabla 5.7. Del mismo modo, la columna cuarta corresponde al vector resultado de *Alta Vista* (VR2) y la quinta columna recoge los valores de relevancia de las URL de este segundo vector.

N	VR1		VR2	
1	**http://www.sivale.com/una.htm**	1	**http://www.sivale.com/una.htm**	1
2	**http://www.sivale.com/dos.htm**	1	**http://www.novale.com/tres.htm**	0
3	http://www.novale.com/una.htm	0	http://tampocovale.com/una.htm	0
4	http://www.novale.com/dos.htm	0	**http://www.sivale.com/dos.htm**	1
5	**http://www.novale.com/tres.htm**	0	*http://estasivale.com/una.htm*	1
6	**http://www.sivale.com/tres.htm**	1	**http://www.sivale.com/tres.htm**	1
7	**http://www.sivale.com/cuatro.htm**	1	http://tampocovale.com/dos.htm	0
8	**http://www.sivale.com/cinco.htm**	1	http://tampocovale.com/tres.htm	0
9	**http://www.sivale.com/seis.htm**	1	http://tampocovale.com/cuatro.htm	0
10	http://www.novale.com/cuatro.htm	0	*http://estasivale.com/dos.htm*	1
11	**http://www.novale.com/cinco.htm**	0	**http://www.sivale.com/cuatro.htm**	0,9
12	http://www.novale.com/seis.htm	0	**http://www.sivale.com/cinco.htm**	0,9
13	http://www.novale.com/siete.htm	0	**http://www.sivale.com/seis.htm**	0,9
14	*http://www.sivale.com/siete.htm*	0,9	**http://www.novale.com/cinco.htm**	0
15	http://www.novale.com/ocho.htm	0	*http://estasivale.com/tres.htm*	0,9
16	**http://www.sivale.com/ocho.htm**	0,9	**http://www.sivale.com/ocho.htm**	0,9
17	**http://www.novale.com/nueve.htm**	0	**http://www.novale.com/nueve.htm**	0
18	*http://www.sivale.com/nueve.htm*	0,9	http://tampocovale.com/cinco.htm	0
19	http://www.novale.com/diez.htm	0	http://tampocovale.com/seis.htm	0
20	*http://www.sivale.com/diez.htm*	0,9	http://tampocovale.com/siete.htm	0

Tabla 5.7 Procedimiento de cálculo de la similitud de dos vectores en nuestro experimento. Tabla de vectores de resultado recuperados y de valores de relevancia asignados a sus URL componentes. Fuente: elaboración propia.

En letra negrita se resaltan las URL del vector de resultados de *Google* (VR1) que también son devueltas por el vector de resultados de *Alta Vista* (VR2). En letra cursiva se resaltan aquellas URL devueltas por un motor pero que no han

sido devueltas por el otro y que han sido consideradas relevantes. Sin resaltar quedan las URL no relevantes que sólo aparecen en un vector de resultados. En el cálculo anterior de similitud se han comparado valores de una distribución unidimensional, pero ahora se dispone de una distribución de elementos con dos atributos: URL y su valor de relevancia ponderada. Para poder calcular la similitud se reduce esta distribución a un espacio n-dimensional común, donde n será el número de URL coincidentes para cada par de vectores de resultados más el número de URL relevantes localizadas en cada motor por separado (a diferencia del ejemplo anterior donde el valor de n es fijo). De esta manera, los iniciales vectores (de naturaleza dual) VR1 y VR2 se transforman en vectores de naturaleza simple V1 y V2, cuyos componentes son los valores de relevancia que presenta cada URL en los motores originales. Esta transformación del espacio vectorial origina la siguiente Tabla 5.8:

BIBLIOTECA VIRTUAL

N	URL	V1	V2	V1*V2	$(V1)^2$	$(V2)^2$
1	http://www.sivale.com/una.htm	1	1	1	1	1
2	http://www.sivale.com/dos.htm	1	1	1	1	1
3	http://www.novale.com/tres.htm	0	0	0	0	0
4	http://www.sivale.com/tres.htm	1	1	1	1	1
5	http://www.sivale.com/cuatro.htm	1	1	1	1	1
6	http://www.sivale.com/cinco.htm	1	1	1	1	1
7	http://www.sivale.com/seis.htm	1	1	1	1	1
8	http://www.novale.com/cinco.htm	0	0	0	0	0
9	http://www.sivale.com/siete.htm	0,9	0	0	0,81	0
10	http://www.sivale.com/ocho.htm	0,9	0,9	0,81	0,81	0,81
11	http://www.novale.com/nueve.htm	0	0	0	0	0
12	http://www.sivale.com/nueve.htm	0,9	0	0	0,81	0
13	http://www.sivale.com/diez.htm	0,9	0	0	0,81	0
14	http://estasivale.com/una.htm	0	1	0	0	1
15	http://estasivale.com/dos.htm	0	1	0	0	1
16	http://estasivale.com/tres.htm	0	0,9	0	0	0,81
			Totales	6.81	9.24	9.62

Tabla 5.8 Procedimiento de cálculo de la similitud de dos vectores en nuestro experimento. Tabla de vectores transformados V1 y V2 y determinación del producto escalar de los vectores y del cuadrado de sus componentes. Fuente: elaboración propia.

Ahora sí se dispone de dos vectores (V1 y V2) a los que aplicar la función de similitud del coseno, para ello se calcula la suma de todos los productos escalares de los componentes de V1 y V2 (6.81) y la suma de los cuadrados de cada uno de los componentes de estos vectores (9.24 y 9.62). A continuación, se determina la raíz cuadrada de estos dos últimos valores (9.42) y se divide por ese valor la suma de los productos escalares (6.81 entre 9.42) obteniéndose un valor de 0,7223, lo que representa que estos dos vectores coinciden en un 72.23 % de sus resultados globales. En la práctica, las URL con valor nulo de relevancia no influyen en los resultados finales, siendo su presencia meramente anecdótica. Realmente, a efectos de calcular la similitud sólo interesan los componentes de relevancia mayor de cero, siempre que deseemos analizar la similitud de la pertinencia de la respuesta de dos motores, como es nuestro caso. En cambio, si el estudio se centra en analizar la similitud de los componentes de las respuestas, entonces habría que modificar el enfoque general y reformar el procedimiento de cálculo. Al trabajar con URL relevantes se obtiene un espacio n-dimensional algo menor que en el caso anterior, y además se simplifican los cálculos, obteniéndose los mismos resultados ya que los productos donde interviene un valor igual a cero son irrelevantes.

n	URL	V1	V2	V1*V2	$(V1)^2$	$(V2)^2$
1	http://www.sivale.com/una.htm	1	1	1	1	1
2	http://www.sivale.com/dos.htm	1	1	1	1	1
3	http://www.sivale.com/tres.htm	1	1	1	1	1
4	http://www.sivale.com/cuatro.htm	1	1	1	1	1
5	http://www.sivale.com/cinco.htm	1	1	1	1	1
6	http://www.sivale.com/seis.htm	1	1	1	1	1
7	http://www.sivale.com/siete.htm	0,9	0	0	0,81	0
8	http://www.sivale.com/ocho.htm	0,9	0,9	0,81	0,81	0,81
9	http://www.sivale.com/nueve.htm	0,9	0	0	0,81	0
10	http://www.sivale.com/diez.htm	0,9	0	0	0,81	0
11	http://estasivale.com/una.htm	0	1	0	0	1
12	http://estasivale.com/dos.htm	0	1	0	0	1
13	http://estasivale.com/tres.htm	0	0,9	0	0	0,81
			Totales	6.81	9.24	9.62

Tabla 5.9 Procedimiento de cálculo de la similitud de dos vectores en nuestro experimento. Vectores transformados V1 y V2 con componentes distintos de cero y determinación del producto escalar y del cuadrado de sus componentes. Fuente: elaboración propia.

Esta modalidad de cálculo de la función de similitud presenta un elemento innovador: en su resultado final influye también la posición que ocupan los documentos relevantes en el resultado.

Es decir, el alineamiento incide en el valor de similitud, tal como sucede también en los cálculos de las medidas E-P. Si en lugar de considerar nuestra función de ponderación de la relevancia, se aplicaran los valores binarios de la relevancia, la tabla anterior se transformaría en:

n	URL	V1	V2	V1*V2	$(V1)^2$	$(V2)^2$
1	http://www.sivale.com/una.htm	1	1	1	1	1
2	http://www.sivale.com/dos.htm	1	1	1	1	1
3	http://www.sivale.com/tres.htm	1	1	1	1	1
4	http://www.sivale.com/cuatro.htm	1	1	1	1	1
5	http://www.sivale.com/cinco.htm	1	1	1	1	1
6	http://www.sivale.com/seis.htm	1	1	1	1	1
7	http://www.sivale.com/siete.htm	1	0	0	1	0
8	http://www.sivale.com/ocho.htm	1	1	1	1	1
9	http://www.sivale.com/nueve.htm	1	0	0	1	0
10	http://www.sivale.com/diez.htm	1	0	0	1	0
11	http://estasivale.com/una.htm	0	1	0	0	1
12	http://estasivale.com/dos.htm	0	1	0	0	1
13	http://estasivale.com/tres.htm	0	1	0	0	0
			Totales	7	10	9

Tabla 5.10 Procedimiento de cálculo de la similitud de dos vectores en nuestro experimento. Vectores transformados V1 y V2 con componentes distintos de cero, valores de relevancia binarios y determinación del producto escalar y del cuadrado de sus componentes. Fuente: elaboración propia.

En este caso, el valor de la similitud sería de 0,7378, ligeramente superior al anterior. Independientemente de la importancia, en términos estrictamente numéricos, de esta diferencia, resulta más apropiado para nuestro estudio medir la similitud bajo un doble prisma: el de la coincidencia de términos y el de la ubicación de estos términos en el vector de resultados, porque ya no se trata de que los documentos relevantes figuren en los vectores de resultado de los distintos motores de búsqueda, sino que también se trata de que al usuario se les muestren de manera más o menos similar.

Fase 7: Agrupamiento de los índices.

La condición indispensable para la realización de esta fase es la necesidad de determinar un comportamiento más o menos similar de los valores de la similitud que se establecen entre dos motores de búsqueda ensayo a ensayo.

Sólo en este caso, cabe la posibilidad de plantear la hipótesis de agrupar los índices de los motores con la idea de ser más efectivos (es decir, más precisos y más exhaustivos), ya que si no estaríamos tomando como base de un agrupamiento circunstancias particulares de similitud y este agrupamiento no sería válido.

La heterogénea naturaleza de la web induce a pensar que esta condición será difícilmente satisfecha por los motores, aunque esta hipótesis debe ser verificada por experimentación suficiente y razonada.

En el caso de que este agrupamiento se vislumbrara posible, a los valores medios de similitud entre los motores se les aplicaría la técnica de agrupamiento conocida como el "vecino más cercano" [PRE, 2000], con el fin de formar los grupos de motores. Posteriormente se calcularían los valores de efectividad de las muestras extraídas de estas agrupaciones de motores y se contrastarían con los resultados obtenidos por separado.

Ejemplo de evaluación.

Se va a proceder a la ejemplificación de la aplicación de esta propuesta de procedimiento de evaluación de la efectividad de los motores de búsqueda con un caso práctico.

En el presente ejemplo se analizan los treinta primeros documentos obtenidos como resultado en seis motores[124] a los que se les pregunta por sitios web relacionados con los alquileres de apartamentos en Málaga o provincia, combinando en las expresiones los operadores booleanos y la búsqueda por frase literal.

En primer lugar se toma nota de los documentos devueltos por cada motor y de los enlaces considerados duplicados y erróneos detectados entre los treinta documentos analizados de la muestra. Los resultados obtenidos en se muestran en la tabla 4.11:

[124] Alta Vista, All the Web, Google, MSN Network, Terra y WISEnut.

	Docs. Localizados	Enlaces erróneos	Enlaces duplicados
Alta Vista	253	0	0
All the Web	258	5	0
Google	568	2	2
MSN	151	1	5
Terra	156	5	8
WISEnut	50	2	0

Tabla 5.11 Ejemplo de aplicación del procedimiento de evaluación de la efectividad. Paso 1: anotación de documentos recuperados e identificación de erróneos y duplicados de las muestras analizadas.

Estos vectores de resultados proporcionan los siguientes valores de *precisión, exhaustividad*, enlaces erróneos y enlaces fallidos.

	Precisión	Exhaustividad	%Enlaces Erróneos	% Enlaces duplicados
Alta Vista	0,848	0,1659	0	0
Alltheweb	0,5342	0,1044	16,66	0
Google	0,9154	0,1791	6,66	6,66
Msn	0,6572	0,1214	3.33	16.66
Terra	0,6051	0,1171	16.66	26.66
WISEnut	0,6509	0,1325	6.66	0

Tabla 5.12 Ejemplo de aplicación del procedimiento de evaluación de la efectividad. Paso 2: cálculo de los pares de valores E-P para cada motor y determinación de porcentajes erróneos y duplicados.

La operación de búsqueda que ha obtenido mejores valores de precisión y exhaustividad, es la realizada por el motor *Google*. El único motor que no se ha visto afectado por la presencia de enlaces erróneos o inactivos ha sido *Alta Vista,* motor que, junto a *All the Web* y *WISEnut* no presenta duplicados.

Como este conjunto de medidas individuales no basta para ofrecer una idea de cuál de los vectores de resultados es de mayor calidad para los usuarios, se va a calcular la medida de valor simple de Borko.

	P	E	I1
Alta Vista	0,848	0,1659	1,0139
Alltheweb	0,5342	0,1044	0,6386
Google	0,9154	0,1791	1,0945
Msn	0,6572	0,1214	0,7786
Terra	0,6051	0,1171	0,7222
WISEnut	0,6509	0,1325	0,7834

Tabla 5.13 Ejemplo de aplicación del procedimiento de evaluación de la efectividad. Paso 3: cálculo de la medida de Borko para cada motor.

La aplicación de esta medida de valor simple confirma que la mejor búsqueda la ha realizado *Google*. Lo que no se está en condiciones de precisar es si el mejor resultado obtenido por *Google* frente a *Alta Vista,* el motor que ha quedado en segundo lugar, tiene algo que ver con el total de documentos recuperados por *Google,* más del doble que *Alta Vista*.

Para poder apreciar esta incidencia resulta necesario *tipificar* la precisión y la exhaustividad con respecto al porcentaje de documentos recuperados por cada motor del universo total recuperado en esta operación de búsqueda.

La cardinalidad de este universo es la suma del total de documentos recuperados por los seis motores objeto de estudio, lo que representa un total de 1436 documentos. Con el fin de atenuar la incidencia de las grandes diferencias presentes en esta medida de tamaño, estos porcentajes se calcularán en términos del logaritmo neperiano de los valores originales. Así, la distribución que a continuación muestra la Tabla 5.14, recoge la distribución proporcional del número de documentos.

	docs	**ln(docs)**	**T**	**|1 – T|**
Avista	253	4,8402	0,1776	0,82
Alltheweb	258	4,8598	0,1783	0,82
Google	568	5,649	0,2073	0,79
Msn	151	4,3241	0,1587	0,84
Terra	156	4,3567	0,1599	0,84
WISEnut	50	3,2189	0,1181	0,88

Tabla 5.14 Ejemplo de aplicación del procedimiento de evaluación de la efectividad. Paso 4: determinación de los porcentajes de documentos recuperados por cada motor (T) sobre el total de documentos recuperados por los seis motores en conjunto.

La idea subyacente en este proceso de normalización de las variables precisión y exhaustividad, es intentar determinar hasta qué grado influye en su cálculo el tamaño del índice del motor de búsqueda.

Así, si en este proceso de tipificación se empleara la distribución del porcentaje que sobre el total posee cada motor (el valor T de la tercera columna de la Tabla 5.14), se volvería a favorecer a los de mayor número de documentos.

Para establecer una distribución relacionada con este porcentaje, pero que penalice (en alguna proporción) a aquellos motores de mayor volumen de documentos en su colección, se pensó en emplear el módulo del complemento a la unidad de este porcentaje T (| **1- T** |) como valor que permitiera introducir esta penalización, ya que obtiene un mejor resultado a medida que el conjunto de documentos recuperado por cada motor es menor.

El siguiente paso de la tipificación es calcular la media aritmética y la desviación típica de los valores de precisión y exhaustividad obtenidos por los seis motores de búsqueda objeto de estudio. Los resultados se muestran en la Tabla 5.15:

	P	**E**
Avista	0,848	0,1659
Alltheweb	0,5342	0,1044
Google	0,9154	0,1791
Msn	0,6572	0,1214
Terra	0,6051	0,1171
WISEnut	0,6509	0,1325
Media	0,7018	0,1367
Desviación	0,1477	0,0294

Tabla 5.15 Ejemplo de aplicación del procedimiento de evaluación de la efectividad. Paso 5: cálculo de la media aritmética y desviación estándar de los pares de valores E-P.

Para determinar la precisión tipificada del motor *Alta Vista*, por ejemplo, se debe calcular la diferencia entre la precisión original de este motor y la media aritmética de las precisiones (0,848 – 0,7018 = 0,1462), y dividir el resultado por la desviación típica de las precisiones (0,1477). El resultado obtenido, que vamos a denominar módulo de la precisión, para este motor será igual a 0.9898.

El cálculo de este módulo de la precisión (y también del módulo de la exhaustividad) es el primer paso del proceso de la tipificación de una variable. El segundo paso es calcular el módulo de la variable **|1 – T|** y el tercero hallar la media aritmética de ambos valores, obteniéndose así la precisión tipificada (y de la misma manera, la exhaustividad tipificada).

Tras realizar varios ensayos, se ha decidido introducir una modificación en este cálculo, ya que ha comprobado que la variable **|1 - T|** incide excesivamente en la determinación de los valores tipificados de exhaustividad y precisión (y por consiguiente, en el cálculo de la medida de valor simple $\mathbf{I_{1\ tipificada}}$), favoreciendo a los motores de menor tamaño frente a los de mayor tamaño, independientemente de que los valores de precisión obtenidos por estos sean óptimos (se ha dado el caso de designar mejor búsqueda la realizada a un motor que apenas sobrepasaba el 50% de precisión, frente a otros que alcanzaban valores cercanos al 90%).

Otra circunstancia observada, algo más intuitiva quizá que la anterior pero igualmente significativa, reside en el hecho de que la distribución de valores de **|1 – T|** es más o menos comparable (en términos de límite superior) a la distribución de la precisión pero, en cambio, no es comparable en absoluto a la distribución de la exhaustividad, ya que los límites superiores de esta segunda variable son muy pequeños en comparación con la distribución de **|1 – T|.**

Por esta serie de razones, se ha considerado oportuno paliar la influencia de esta variable tipificada dividiendo sus valores entre dos a la hora de calcular las medida tipificadas de la precisión y de la exhaustividad, reduciendo así su influencia y propiciando que su distribución se encuentre igual de próxima a la distribución de los valores de la precisión que de la exhaustividad.

Siguiendo con el ejemplo de la precisión tipificada de Alta Vista, se determinan los valores de la variable **|1 – T|** para este motor, obteniéndose un porcentaje tipificado negativo de -0.39. Finalmente se suma este valor (dividido entre dos) con el valor de la precisión y se halla la media aritmética, obteniéndose el valor de la precisión normalizada del motor *Alta Vista,* aproximadamente 0.3.

Similar procedimiento se debe seguir con el resto de los motores para calcular el resto de valores de la precisión y de la exhaustividad tipificadas (que pueden consultarse en la Tabla 5.15).

	P	E	docs	PN	EN
Avista	0,848	0,1659	253	0,39	0,39
Alltheweb	0,5342	0,1044	258	-0,66	-0,64
Google	0,9154	0,1791	568	0,37	0,37
Msn	0,6572	0,1214	151	-0,08	-0,19
Terra	0,6051	0,1171	156	-0,25	-0,26
WISEnut	0,6509	0,1325	50	0,23	0,33
Media	0,7018	0,1367			
Desviación	0,1477	0,0294			

Tabla 5.16 Ejemplo de aplicación del procedimiento de evaluación de la efectividad. Paso 6: cálculo de la precisión y de la exhaustividad tipificadas.

El siguiente paso es calcular de nuevo la medida I_1 de Borko, que en este caso, al partir de datos tipificados, sería la medida $I_{1\ tipificada}$, cuyos valores se recogen en la Tabla 5.17:

	PN	EN	I1	I1tipificada
Avista	3,21	2,96	1,0139	**0,78**
Alltheweb	2,1	2,25	0,6386	-1,3
Google	3	2,67	1,0945	0,74
Msn	2,66	2,66	0,7786	-0,27
Terra	2,56	2,03	0,7222	-0,51
WISEnut	3,3	2,89	0,7834	0,56

Tabla 5.17 Ejemplo de aplicación del procedimiento de evaluación de la efectividad. Paso 7: cálculo de la medida I1 tipificada.

Según estos nuevos cálculos de las medidas de valor simple, la búsqueda realizada por *Alta Vista* es ligeramente mejor que la realizada por *Google* (que, en principio, se veía muy favorecido por poseer un índice de mayor tamaño).

Otro aspecto interesante de esta medida tipificada es la penalización que sufren los valores de los motores *All the Web* y *Terra*, el primero por sus bajos valores de precisión y el segundo por no haber presentado valores altos de exhaustividad. Queda por presentar el método de agrupación de los vectores de resultado con base en su similitud. La siguiente tabla recoge los

valores que ofrece la función del coseno para cuatro de los motores analizados[125].

	All the Web	***Google***	***MSN***
Alta Vista	0,26	0,28	0,32
All the Web		0,50	0,13
Google			0,15

Tabla 5.18 Ejemplo de aplicación del algoritmo de alimenamiento por el método del "Vecino más Cercano". Paso 0: tabla de similitudes medias.

La similitud media de los vectores de resultado es 0,27. Suponiendo que estos valores de similitud media son los que representan las medidas centrales de esta función, podemos determinar a partir de ellos y aplicando la técnica de agrupamiento conocida como "el vecino más cercano", los siguientes agrupamientos:

Paso 1: se parte de la tabla de similitudes calculada y se transforman las similitudes en distancias (restando a la unidad el valor de la similitud).

	All the Web	***Google***	***MSN***
Alta Vista	0,74	0,72	0,68
All the Web		0,50	0,87
Google			0,85

Tabla 5.19 Ejemplo de aplicación del algoritmo de alimenamiento por el método del "Vecino más Cercano". Paso 1: tabla de distancias medias.

Paso 2: Se identifica la menor de las distancias medias y se lleva a cabo el primero de los agrupamientos entre los distintos motores.

[125] *Alta Vista*, *All the Web*, *Google* y *Microsoft Network*.

	Alta Vista	***MSN***
All the Web – Google	0.72	0.85
Alta Vista	0,74	0,68

Tabla 5.20 Ejemplo de aplicación del algoritmo de alimenamiento por el método del "Vecino más Cercano". Paso 1: establecimiento del primer agrupamiento.

El valor resultante de las distancias medias de los motores Alta Vista y MSN con el nuevo agrupamiento, será el menor de los valores de distancia que mantenía en un inicio con cada uno de ellos de forma individual.

Paso 3: Al igual que en el paso anterior, se identifica la menor de las distancias medias y se lleva a cabo el segundo de los agrupamientos entre los distintos motores.

	Alta Vista – MSN
All the Web – Google	0.72
Alta Vista - MSN	

Tabla 5.21 Ejemplo de aplicación del algoritmo de alimenamiento por el método del "Vecino más Cercano". Paso 3: establecimiento del segundo agrupamiento.

Paso 4: El último paso correspondería a establecer el agrupamiento entre los dos subconjuntos de motores identificados. Así, resultaría que los cuatro motores de este ejemplo se agruparían a una distancia de 0.72, valor relativamente alto para identificar subgrupos de motores dentro de esta colección claramente identificados, más si se tiene en cuenta que el primero de los agrupamientos se ha realizado a una distancia de 0.50 (que equivale a decir que la mitad de los índices de los motores son absolutamente diferentes).

No obstante, este método, si los resultados que se obtuvieran en el estudio así lo permitieran, favorecería la identificación de los motores más afines, siempre que las distancias medias establecidas entre los componentes de estos grupos más próximos no alcanzaran valores tan altos como en este ejemplo.

Tablas e ilustraciones.

Tabla 5.1 Cuestiones empleadas para medir la distribución porcentual de los distintos motores con base en una serie de preguntas en idioma Español. Fuente: Elaboración propia. 178

Ilustración 5.1 Distribución porcentual de los distintos motores con base en una serie de preguntas en idioma Español. Fuente: Elaboración propia.....179

Tabla 5.2 Sintaxis empleada en Google en el experimento de análisis de la efectividad de la recuperación de información según el tipo de búsqueda. Fuente: Elaboración propia. ...182

Tabla 5.3 Pares de valores E-P obtenidos en el experimento destinado a analizar la efectividad de la recuperación de información según el tipo de búsqueda empleado. Fuente: Elaboración propia. 183

Tabla 5.4 Formulario empleado para indicar a priori el objeto de cada búsqueda del experimento. Fuente: Elaboración propia.....184

Tabla 5.5 Función de ponderación del grado de relevancia de un documento en función de su alineamiento. Fuente: Elaboración propia.....186

Ilustración 5.2 Sección del resultado de una búsqueda realizada en el motor Google. Se muestran dos recursos del mismo sitio web, uno específico y otro más general. Fuente: Elaboración propia. 187

Ilustración 5.6 Fórmula de la función de similitud del Coseno. Fuente: 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 193

Tabla 5.6 Cálculo de la función de similitud del Coseno para el ejemplo de los vectores P y N. Fuente: Elaboración propia.....193

Ilustración 5.7 Significado de la función de similitud del coseno. Fuente: 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 194

Tabla 5.7 Procedimiento de cálculo de la similitud de dos vectores en nuestro experimento. Tabla de vectores de resultado recuperados y de valores de relevancia asignados a sus URL componentes. Fuente: elaboración propia. 195

Tabla 5.8 Procedimiento de cálculo de la similitud de dos vectores en nuestro experimento. Tabla de vectores transformados V1 y V2 y determinación del producto escalar de los vectores y del cuadrado de sus componentes. Fuente: elaboración propia. 196

Tabla 5.9 Procedimiento de cálculo de la similitud de dos vectores en nuestro experimento. Vectores transformados V1 y V2 con componentes distintos de cero y determinación del producto escalar y del cuadrado de sus componentes. Fuente: elaboración propia. 197

Tabla 5.10 Procedimiento de cálculo de la similitud de dos vectores en nuestro experimento. Vectores transformados V1 y V2 con componentes distintos de cero, valores de relevancia binarios y determinación del producto escalar y del cuadrado de sus componentes. Fuente: elaboración propia.....198

Tabla 5.11 Ejemplo de aplicación del procedimiento de evaluación de la efectividad. Paso 1: anotación de documentos recuperados e identificación de erróneos y duplicados de las muestras analizadas. 200

Tabla 5.12 Ejemplo de aplicación del procedimiento de evaluación de la efectividad. Paso 2: cálculo de los pares de valores E-P para cada motor y determinación de porcentajes erróneos y duplicados. 200

Tabla 5.13 Ejemplo de aplicación del procedimiento de evaluación de la efectividad. Paso 3: cálculo de la medida de Borko para cada motor.....201

Tabla 5.14 Ejemplo de aplicación del procedimiento de evaluación de la efectividad. Paso 4: determinación de los porcentajes de documentos recuperados por cada motor (T) sobre el total de documentos recuperados por los seis motores en conjunto. 201

Tabla 5.15 Ejemplo de aplicación del procedimiento de evaluación de la efectividad. Paso 5: cálculo de la media aritmética y desviación estándar de los pares de valores E-P. 202

Tabla 5.16 Ejemplo de aplicación del procedimiento de evaluación de la efectividad. Paso 6: cálculo de la precisión y de la exhaustividad tipificadas. 204

Tabla 5.17 Ejemplo de aplicación del procedimiento de evaluación de la efectividad. Paso 7: cálculo de la medida I1 tipificada.....204

Tabla 5.18 Ejemplo de aplicación del algoritmo de alimenamiento por el método del "Vecino más Cercano". Paso 0: tabla de similitudes medias. 205

Tabla 5.19 Ejemplo de aplicación del algoritmo de alimenamiento por el método del "Vecino más Cercano". Paso 1: tabla de distancias medias. 205

Tabla 5.20 Ejemplo de aplicación del algoritmo de alimenamiento por el método del "Vecino más Cercano". Paso 1: establecimiento del primer agrupamiento.....206

Tabla 5.21 Ejemplo de aplicación del algoritmo de alimenamiento por el método del "Vecino más Cercano". Paso 3: establecimiento del segundo agrupamiento. 206

BIBLIOTECA VIRTUAL DE MIGUEL DE CERVANTES

6 Aplicación de la propuesta de modelo de evaluación de los SRI en la web.

RESUMEN: Este capítulo recoge el ensayo realizado sobre seis motores de búsqueda con el objeto de verificar la viabilidad de nuestra propuesta de modelo muldimensional de evaluación del funcionamiento de estos sistemas Los motores elegidos han sido *Alta Vista*, *All the Web*, *Google*, *MSN* Network, *Terra* y *WISEnut*, seleccionados de una muestra mayor de sistemas a los que se les exigió una serie de características mínimas para formar parte de este ensayo. Los resultados de este ensayo demuestran la viabilidad de la propuesta, ya que dimensión a dimensión analizada resultan coincidentes con la mayoría de estudios realizados con anterioridad y permiten el establecimiento, desde un punto de vista inductivo, de un funcionamiento tipo de los motores de búsqueda.

Aplicación de la propuesta de modelo de evaluación de los SRI en la web.

Elección del conjunto de motores. 211

Temática de las búsquedas....215

Generalidad de los motores de búsqueda....215
Relación de cuestiones realizadas....215

Datos relacionados con la *Precisión*....219

Precisión de las operaciones con intersección. 219
Precisión de las operaciones con intersección y frase literal. 224
Comparativa de las medias de los valores obtenidos....227

Datos relacionados con la *Exhaustividad* 227

Exhaustividad de las operaciones con intersección. 228
Exhaustividad de las operaciones con intersección y frase literal. 230

Datos sobre total de documentos localizados. 231

Datos sobre enlaces erróneos y duplicados. 233

Datos sobre enlaces erróneos....233
Datos sobre enlaces duplicados. 236

Datos sobre la similitud de los resultados. 239

Medias de las similitudes obtenidas. 239
Medias de las similitudes obtenidas motor a motor. 240
Similitud de resultados obtenidos por los motores con ambas modalidades de búsqueda en el mismo tema. 243

Datos sobre la medida de valor simple de Borko. 244

Datos sobre la medida de valor simple de Borko tipificada. 247

Datos sobre los agrupamientos de los motores. 250

Agrupamientos con muestras de diez documentos. 251
Agrupamientos con muestras de veinte documentos. 252
Agrupamientos con muestras de treinta documentos. 254

Reflexión general sobre el comportamiento de los motores de búsqueda analizados.256

Tablas e ilustraciones. 258

Elección del conjunto de motores.

La elección del conjunto de motores se realiza cumpliendo las directrices de la propuesta de evaluación. Como las búsquedas emplearán dos tipos de operaciones (intersección simple e intersección combinada con búsqueda por frase literal), en idioma Español y sin restricciones geográficas, los motores deberán cumplir los siguientes requisitos[126]:

1. Búsqueda por intersección de términos.
2. Búsqueda por frase literal.
3. Combinación de las dos búsquedas anteriores.
4. Especificación del idioma de los documentos.
5. Ausencia de sesgos en el alcance geográfico de las búsquedas.
6. Indicación del número de documentos recuperados[127]

La web *Searchengine Watch* muestra la selección de motores [SEA, 2002], que sirve de base para el establecimiento de la muestra de motores. A partir de esta documentación, se construye la tabla 6.1:

		Requisitos					
Motor	**URL**	**1**	**2**	**3**	**4**	**5**	**6**
Alltheweb	http://www.alltheweb.com	X	X	X	X	X	X
Alta Vista	http://es-es.altavista.com	X	X	X	X	X	X
AOL Search	http://www.aol.com	X	X	X		X	X
Ask Jeeves	http://www.askjeeves.com		X			X	
Direct Hit	http://www.directhit.com	X	X	X		X	
Google	http://www.*Google*.com	X	X	X	X	X	X
Hot Bot	http://www.hotbot.com	X	X	X	X	X	X
Iwon	http://home.iwon.com	X	X	X		X	X
Inktomi	http://www.inktomi.com/	X	X	X	X	X	X
Lycos	http://www.lycos.com/	X	X	X	X	X	X
MSN	http://search.*MSN*.com/	X	X	X	X	X	X

Tabla 6.1 Requisitos de la propuesta de evaluación satisfechos por cada motor de búsqueda. Fuente: elaboración propia.

[126] El motor debería permitir usar operadores booleanos.

[127] Vital para el cálculo de los valores E-P de las búsquedas realizadas en cada motor. Sorprendentemente, algunos motores no informan de este dato.

La tabla 6.1 recoge el grado de concordancia existente entre las prestaciones de cada motor y los requisitos indicados anteriormente[128]. Del conjunto inicial de motores se eligen aquellos que cumplan todos los requisitos, es decir: Alltheweb, *Alta Vista*, *Google*, *Hot Bot*, *Inktomi*, *Lycos* y *MSN*. En este primer subconjunto de motores candidatos a formar parte de nuestro estudio se presentan las siguientes incidencias:

- *Alltheweb*, *Hot Bot* y *Lycos* comparten la misma tecnología de búsqueda y además los resultados de varias búsquedas pilotos que hemos realizado nos inducen a pensar que comparten también la colección de documentos. Es por ello que sólo vamos a elegir uno de los tres para formar parte del experimento, con el fin de no desvirtuar los valores de similitud que se obtengan (en este caso serían más altos de lo normal[129]). Por la sencillez de su interface de búsqueda nos hemos decantado por *Alltheweb*.
- *Inktomi* realmente es un motor de búsqueda que emplean diversos servicios de información en Internet. Si uno se conecta a la página principal de la web de *Inktomi*[130], no va a encontrar un formulario de búsqueda para recuperar distintos recursos en la web. Lo más parecido que encuentra es un enlace a un formulario para sugerir una dirección de Internet que será recogida posteriormente en servicios como *MSN* o *AOL*[131]. Esta empresa pone su colección de documentos y su tecnología al servicio de otros portales de Internet, tales como los dos citados anteriormente y el portal español *Terra Networks* (*Terra* en adelante). Es por ello que no podemos evaluar el motor *Inktomi*, sino algunos de los servicios de información que hacen uso del mismo. En nuestro caso, vamos a elegir *MSN* y *Terra*, que cumplen todos los requisitos especificados.
- *AOL* no permite especificar el idioma de los documentos a recuperar, por lo cual va a quedar excluido de nuestro estudio.

---

[128] En la construcción de esta tabla se excluyen los directorios que aparecen recogidos en ese informe: *Yahoo, Netscape, Open Directory* y *Looksmart*, entre otros.

[129] De hecho, en las primeras ocho búsquedas, de las treinta que han formado parte de nuestro experimento, incluimos el motor *Lycos* y los valores de similitud con el motor *Alltheweb* eran cercanos a la unidad, muy alejados de la media de similitudes del resto de motores.

[130] http://www.inktomi.com

[131] AOL son las siglas del portal *American OnLine*.

La elección de *MSN* y *Terra* nos permitirá también analizar hasta qué punto comparten información y hasta qué punto, cada motor ha desarrollado su colección con las aportaciones de los usuarios de los servicios y el funcionamiento por separado de sus agentes de búsqueda[132].

Así, tras esta serie de consideraciones, el subconjunto de motores candidatos para nuestro estudio es el siguiente: *Alltheweb, Alta Vista, Google*, *MSN* y *Terra*. A este subconjunto, nuestra idea original consistía en añadir otros dos motores (no recogidos en la web *Searchengine Watch*): *WISEnut* y *Northernlight*. Con estos dos motores se totalizarían siete. La revolución que ha representado *Google* en este campo, al considerar el número de enlaces de una página como elemento determinante del alineamiento de las direcciones devueltas en una operación de búsqueda, no se ha detenido exclusivamente en este motor, sino que está siendo aplicado en otros sistemas. Tal es el caso de nuevos motores que se vienen desarrollando desde hace relativamente poco tiempo, tales como *WISEnut*, *Teoma*, *Lasoo* y *Vivisimo* [GON, 2001].

De este conjunto, el que posee una significativa base de datos (de volumen muy superior al resto de los demás) es el primero de ellos, *WISEnut (con un* índice de documentos en idioma Español de dimensiones parecidas a los índices de *Terra* o *MSN*, tal como se ha podido comprobar en el experimento desarrollado a tal efecto en el capítulo anterior).

La introducción de este motor resulta especialmente interesante por otro motivo, permitirá contrastar el rendimiento de un algoritmo de extensión de los enlaces en dos colecciones de documentos diferentes: la de *Google* y la de *WISEnut*.

El segundo motor que, en un inicio, figuraba en la colección objeto de nuestro estudio, es el motor *Northern Light*. El criterio que propiciaba esta elección se basaba en que el mismo había formado parte de varios estudios recogidos en el capítulo tercero, presentando un rendimiento interesante en las evaluaciones a las que ha sido sometido, además de poseer una colección de un tamaño considerable, tal como se ha visto también en varios estudios recogidos en ese mismo capítulo. De hecho, este motor formó parte del estudio durante varios meses pero, cuando se habían realizado aproximadamente el sesenta por ciento de las búsquedas necesarias para nuestra experimentación, este motor dejó de ser de libre acceso, pasando a

---

[132] Se ha comentado en el capítulo tercero que el motor *Inktomi* dispone de dos colecciones de documentos: "Lo mejor del web", que agrupa a 110 millones de documentos y "El resto del web", que agrupa 390 millones de páginas [SUL, 2001a].

formar parte de la oferta de servicios de pago que la empresa propietaria, *Divine Inc*. proporciona a sus clientes.

El acceso libre ha quedado restringido a unas colecciones de muy reducido tamaño en nada comparables con la colección entera del motor. Con el fin de poder finalizar el estudio con la presencia de este motor se entablaron gestiones con los representantes de esta empresa, quienes declinaron facilitar el acceso a la colección completa con la excusa de que este sistema no podía compararse con los motores de propósito general. Por ello, a pesar de haber realizado una considerable cantidad de cálculos con este motor, se tuvo que excluir del análisis. En último lugar, se deben mencionar otros motores, "quizás algo menos populares y con menos documentos en sus colecciones que los anteriores, pero que gozan también de cierto reconocimiento en la web" [SEA, 2001]. De este grupo de motores "de segunda línea", destacan *Goto* y *Excite*.

El primero de ellos no poseía la posibilidad de realizar operaciones de intersección entre los términos integrantes de la ecuación de búsqueda, ni tampoco permitía la búsqueda literal, por lo cual no podía formar parte de nuestro experimento. El segundo de ellos, *Excite*, a pesar de su pobre interface de usuario, sí permitía este tipo de búsqueda y también formó parte inicialmente del conjunto de motores analizados en nuestro estudio, aunque fue excluido con posterioridad al ser absorbido (junto con *Goto*) por el nuevo motor *Overture*[133], que posee la misma tecnología de recuperación de información que *Goto* y por tanto no puede formar parte de nuestro estudio. Tras toda esta serie de consideraciones, elegimos como muestra objeto de nuestro proceso de análisis el siguiente conjunto de motores de búsqueda:

1. *Alltheweb*
2. *Alta Vista*
3. *Google*
4. *MSN*
5. *Terra*
6. *WISEnut*

[133] *Overture* también ha absorbido a otro conocido motor recogido en varios estudios anteriores, *Infoseek*. Todo este proceso de fusiones entre motores se ha producido de forma paralela a nuestro estudio, por lo que hemos tenido que ir adaptando la muestra.

Temática de las búsquedas.

En la realización de nuestro experimento se ha pretendido diversificar ampliamente las temáticas de las búsquedas, con el fin de intentar reflejar al máximo la heterogénea realidad del uso de estos sistemas en la web.

Generalidad de los motores de búsqueda.

Los motores de búsqueda no son sólo empleados por usuarios expertos en una materia que buscan profundizar en sus conocimientos, o bien pretenden hallar puntos de vista diferentes sobre la materia objeto de su especialidad, estos sistemas de recuperación de información también son empleados para encontrar academias de idiomas en una ciudad, para conseguir teléfonos de inmobiliarias para alquilar un apartamento en vacaciones, para hallar la dirección de una *Librería de Antiguo* donde comprar un ejemplar original de la obra "Breves nociones de Historia de España" de Bernardo Álvarez y Marina, para localizar sitios de Internet donde comprar billetes de avión directamente sin tener que pasar por la agencia de viajes o para que un estudiante de *Educación Secundaria* encuentre la biografía de *Isaac Newton* o de *Charles Darwin,* con el fin de completar un trabajo de clase.

Como los motivos del uso de estos sistemas de recuperación de la información pueden ser muy variados, la temática de preguntas que formará parte del estudio también ha de ser amplia y diversa, propiciándose así un mayor acercamiento de nuestro estudio al contexto de la web y procurando proporcionar una serie de resultados más cercanos al usuario normal que al usuario especializado.

Relación de cuestiones realizadas.

Se han realizado un total de treinta preguntas en cada uno de los motores. Los temas han sido los recogidos en la Tabla 6.2.a. Se puede comprobar que las preguntas poseen ya un cierto grado de formalización, en referencia a las hipotéticas necesidades de información de un usuario particular.

El proceso de transformación de las necesidades de información del usuario en una vista lógica (o "query") para la búsqueda en el SRI, es una etapa de la recuperación de la información que queda fuera de este estudio.

1. Alquiler de apartamentos en Málaga
2. Compra-Venta de automóviles de segunda mano en Madrid
3. Recetas de cocina y dieta mediterránea
4. Librerías de antiguo en España
5. Empresas de fabricación de calzado en Alicante
6. Academias de idiomas en Valencia
7. Galerías de Arte en Murcia
8. Apuntes de Estadística Descriptiva
9. Diseño de sistemas multimedia para el aprendizaje
10. Evaluación de la calidad de la enseñanza universitaria
11. Estrategias de Representación del Conocimiento
12. Diseño accesible a páginas web
13. Historia del Camino de Santiago
14. Temario de Oposiciones de Matemáticas en Secundaria
15. Influencia de la televisión en los niños
16. Modelos pedagógicos para la educación a distancia
17. Principio de incertidumbre de Heisenberg
18. Historia de la ciudad de Ceuta
19. Apuntes de Sistemas Digitales
20. Turismo rural en la Sierra del Segura
21. Apuntes de Historia del Arte Barroco
22. Recopilación de Legislación en Derecho Civil
23. Curso a distancia de Programación en PHP
24. Principio de Conservación de la Energía
25. Teoría de la Evolución de Darwin
26. Discurso del Método de Descartes
27. Semana Santa en Murcia
28. Literatura Española en el Siglo de Oro
29. Plan de Estudios de Licenciado en Comunicación Audiovisual
30. Bibliografía de Miguel de Unamuno

Tabla 6.2.a Temas de las treinta preguntas realizadas a los motores de búsqueda seleccionados. Fuente: elaboración propia.

La Tabla 6.2.b recoge las expresiones formales que se han empleado para la realización de las treinta preguntas con el operador de la intersección:

1. Alquiler AND apartamentos AND Málaga
2. Compra AND Venta AND automóviles AND segunda AND mano AND Madrid
3. Recetas AND cocina AND dieta AND mediterránea
4. Librerías AND antiguo AND España
5. Empresas AND fabricación AND calzado AND Alicante
6. Academias AND idiomas AND Valencia
7. Galerías AND Arte AND Murcia
8. Apuntes AND Estadística AND Descriptiva
9. Diseño AND sistemas AND multimedia AND aprendizaje
10. Evaluación AND calidad AND enseñanza AND universitaria
11. Estrategias AND Representación AND Conocimiento
12. Diseño AND accesible AND páginas AND web
13. Historia AND Camino AND Santiago
14. Temario AND Oposiciones AND Matemáticas AND Secundaria
15. Influencia AND televisión AND niños
16. Modelos AND pedagógicos AND educación AND distancia
17. Principio AND incertidumbre AND Heisenberg
18. Historia AND ciudad AND Ceuta
19. Apuntes AND Sistemas AND Digitales
20. Turismo AND rural AND Sierra AND Segura
21. Apuntes AND Historia AND Arte AND Barroco
22. Recopilación AND Legislación AND Derecho AND Civil
23. Curso AND distancia AND Programación AND PHP
24. Principio AN D Conservación AND Energía
25. Teoría AND Evolución AND Darwin
26. Discurso AND Método AND Descartes
27. Semana AND Santa AND Murcia
28. Literatura AND Española AND Siglo AND Oro
29. Plan AND Estudios AND Licenciado AND Comunicación AND Audiovisual
30. Bibliografía AND Miguel AND Unamuno

Tabla 6.2.B Expresiones formales de las treinta preguntas realizadas a los motores de búsqueda seleccionados con el operador de la intersección. Fuente: elaboración propia.

La Tabla 6.2.c recoge las expresiones formales que se han empleado para la realización de las treinta preguntas combinadamente la operación de intersección y la búsqueda por frase literal.

1. "Alquiler de apartamentos" AND Málaga
2. Compra AND Venta AND automóviles AND "segunda mano" AND Madrid
3. "Recetas de cocina" AND "dieta mediterránea"
4. "Librerías de antiguo" AND España
5. Empresas AND "fabricación de calzado" AND Alicante
6. "Academias de idiomas" AND Valencia
7. "Galerías de Arte" AND Murcia
8. Apuntes AND "Estadística Descriptiva"
9. Diseño AND "sistemas multimedia" AND aprendizaje
10. "Evaluación de la calidad" AND "enseñanza universitaria"
11. Estrategias AND "Representación AND Conocimiento"
12. "Diseño accesible" AND "páginas web"
13. Historia AND "Camino de Santiago"
14. "Temario de Oposiciones" AND Matemáticas AND Secundaria
15. "Influencia de la televisión" AND niños
16. "Modelos pedagógicos" AND "educación a distancia"
17. "Principio de incertidumbre" AND Heisenberg
18. "Historia de la ciudad" AND Ceuta
19. Apuntes AND "Sistemas Digitales"
20. "Turismo rural" AND "Sierra del Segura"
21. Apuntes AND "Historia del Arte" AND Barroco
22. "Recopilación de Legislación" AND "Derecho Civil"
23. "Curso a distancia" AND "Programación en PHP"
24. "Principio de Conservación" AND Energía
25. "Teoría de la Evolución" AND Darwin
26. "Discurso del Método" AND Descartes
27. "Semana Santa" AND Murcia
28. "Literatura Española" AND "Siglo de Oro"
29. "Plan de Estudios" AND Licenciado en Comunicación Audiovisual"
30. Bibliografía AND "Miguel de Unamuno"

Tabla 6.2.c Expresiones formales de las treinta preguntas realizadas a los motores de búsqueda seleccionados con el operador de la intersección y la búsqueda por frase literal. Fuente: elaboración propia.

Datos relacionados con la *Precisión*.

Una vez analizadas y evaluadas todas las direcciones de Internet correspondientes a los treinta primeros resultados devueltos por cada motor de búsqueda, se han obtenido los valores de *precisión* y *exhaustividad* relativos, aplicándose para ello la fórmula que nos presentan Salton y Mc Gill [SAL, 1983].

Se ha calculado la *precisión* para el conjunto de búsquedas que han empleado la operación de intersección, y para el conjunto de búsquedas que han empleado de forma conjunta las operaciones de intersección y frase literal. También se ha calculado una comparativa entre los valores medios de *precisión* obtenidos por cada motor en cada una de estas modalidades de búsquedas.

Los datos correspondientes al cálculo de los valores de *precisión* para cada motor y pregunta están recogidos en el Anexo (apartado "Cálculo de *Precisión*").

***Precisión* de las operaciones con intersección.**

Los resultados obtenidos en este análisis confirman lo apuntado por algunos estudios anteriores (especialmente el realizado por Gordon y Pathak [GOR, 1999]), relativo al comportamiento más o menos similar de este parámetro en los distintos motores de búsqueda, ya que, en la mayor parte de ellos no se aprecian diferencias que puedan considerarse significativas.

De todos modos, tal como se aprecia fácilmente en la Tabla 6.3, los motores *Google* y *Alta Vista* marcan una cierta distancia del resto de los motores, el primero de ellos por sus buenos resultados y el otro por los discretos resultados obtenidos.

	10 docs	**20 docs**	**30 docs**
Altavista	0,43	0,39	0,36
Alltheweb	0,58	0,54	0,51
Google	0,78	0,72	0,68
MSN	0,68	0,61	0,56
Terra	0,63	0,59	0,53
WISEnut	0,54	0,49	0,45

Tabla 6.3 Precisiones medias obtenidas por los motores de búsqueda con la operación de la intersección tras la realización de las 30 preguntas. Fuente: elaboración propia.

Antes de continuar resulta conveniente tener en cuenta que a lo largo de un estudio de naturaleza estadística, "pueden presentarse en la práctica situaciones en las que exista una teoría preconcebida relativa a la característica de la población sometida a estudio. Tal sería el caso, por ejemplo, si pensamos que un tratamiento nuevo puede tener un porcentaje de mejoría mayor que otro estándar, o cuando nos planteamos si los niños de las distintas comunidades españolas tienen la misma altura" [UDB, 2000].

Este tipo de circunstancias son las que propician el estudio de la parcela de la Estadística Inferencial que se recoge bajo el título genérico de *Contraste de Hipótesis*, que implica, en cualquier investigación, la existencia de dos teorías o hipótesis implícitas (hipótesis nula e hipótesis alternativa), que de alguna manera reflejarán esa idea a priori que se posee y que se pretende contrastar con la "realidad".

Este tipo de contrastes están concebidos para evitar la "aparición, implícita de diferentes tipos de errores que podemos cometer durante el procedimiento. No podemos olvidar que, habitualmente, el estudio y las conclusiones que obtengamos para una población cualquiera, se habrán apoyado exclusivamente en el análisis de sólo una parte de ésta" [UDB, 2000].

En nuestro estudio se manifiesta de forma perentoria la necesidad de llevar a cabo estos contrastes con todas las medias que obtengamos de los resultados de cada motor de búsqueda, en tanto que, por muy grande que llegue a ser nuestra muestra, nunca va a poder representar un porcentaje elevado en comparación con el tamaño del índice del motor de búsqueda.

Es por ello que debemos asegurarnos que las medias obtenidas tras la realización de nuestro experimento (como las recogidas en la tabla anterior) son suficientemente representativas.

Es por ello vamos que se va a hacer uso del software de análisis estadístico SPSS[134], por medio del cual aplicaremos un test de contraste paramétrico para cada valor de media aritmética obtenida, siguiendo el ejemplo que nos presenta Marín Fernández en su "Manual de Prácticas" [MAR, 2001a].

Los resultados de la aplicación de este test a todas las medias forman parte del Anexo (apartado "Contrastes"). A continuación vamos a mostrar el procedimiento seguido para verificar la validez de los valores medios obtenidos para el motor *Alta Vista*, con un total de diez documentos analizados.

---

[134] Marca propiedad de SPSS Inc.

Contraste de la media de *Alta Vista* con 10 documentos

Paso 1.

Se establece la hipótesis nula, asignándole a H_0 el valor 0,43; por lo tanto, la hipótesis alternativa H_1 negaría esta afirmación.

Paso 2.

Se obtienen los estadísticos gracias al software SPSS (Menú de Análisis, Comparación de Medias mediante la prueba T de Student), de la distribución de valores de *precisión* obtenidos por *Alta Vista* en los diez primeros documentos de cada pregunta.

(0,8904 - 0,5206 - 0,8471 - 0,4999 - 0,5414 - 0,387 - 0,4024 - 0,8257 - 0,5169 - 0,3028 - 0 - 0,4103 - 0,4899 - 0,3003 - 0,1491 - 0,2744 - 0,5839 - 0,0645 - 0,2574 - 0,01 - 0,6386 - 0,4124 - 0,4453 - 0,4857 - 0,5446 - 0,286 - 0,7992 - 0,5299 - 0,0647 - 0,505)

Estadísticos para una muestra

	N	Media	Desviación típ.	Error típ. de la media
PDOC10	30	,4328	,2373	4,332E-02

Prueba para una muestra

	Valor de prueba = 0.43					
					Intervalo de confianza para la diferencia	
	t	gl	Sig. (bilateral)	Diferencia de medias	Inferior	Superior
PDOC10	,066	29	,948	2,847E-03	-8,58E-02	9,145E-02

Como el valor obtenido en este test del estadístico *Sig. (bilateral)* es 0,948, mucho mayor que 0,05, adoptamos como válida la hipótesis nula H_0 y por lo tanto, la media aritmética de la *precisión* de *Alta Vista* es 0,43

Recuadro 6.1 Ejemplo de Contraste de la media obtenida para *Alta Vista* con 10 documentos analizados. Fuente: elaboración propia.

Siguiendo similar procedimiento para el resto de las medias obtenidas, tanto por *Alta Vista* como por el resto de los motores, se ha verificado la representatividad de las medidas recogidas en la anterior Tabla 6.3.

En la misma se observa expeditamente que todos los motores de búsqueda alcanzan mejores resultados de *precisión* cuando se analizan los diez primeros documentos recuperados, descendiendo de forma progresiva los valores obtenidos cuando aumenta el tamaño de la muestra analizada.

Aunque con unas ligeras diferencias, este descenso de niveles de *precisión* es similar en todos los motores analizados. En la Ilustración 6.1 se observa mejor este comportamiento.

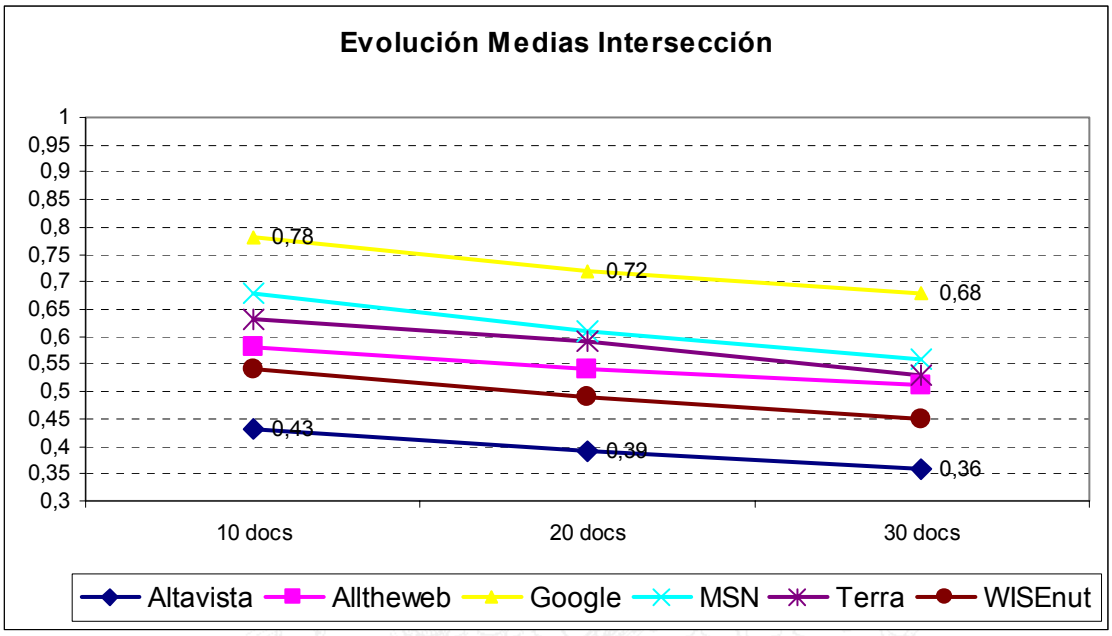


Ilustración 6.1 Evolución de las precisiones medias de todos los motores con la operación de intersección. Fuente: elaboración propia.

Los mejores resultados los ha obtenido *Google*, en un rango comprendido entre el 78% (al analizar diez documentos) y el 68% (al analizar treinta documentos), sólo el motor *MSN* está cercano a estos niveles (cuando se analizan los diez primeros documentos).

También se ha estudiado el *Coeficiente de Variación de Pearson* [QUE, 1982], [AUL, 2000], de las distribuciones de resultados de los distintos motores de búsqueda.

Este coeficiente se obtiene a partir del cociente de la *desviación típica* entre la *media aritmética de la distribución* y permite estudiar la distribución de los valores de la serie, analizando si estos se encuentran más o menos concentrados, o más o menos dispersos. Los resultados se presentan en la Tabla 6.4:

	10 docs	**20 docs**	**30 docs**
Altavista	0,5481	0,5206	0,5074
Alltheweb	0,4553	0,3634	0,3609
Google	0,2608	0,2320	0,2260
MSN	0,3761	0,3482	0,3484
Terra	0,4190	0,4036	0,4423
WISEnut	0,5441	0,4902	0,4468

Tabla 6.4 Coeficiente de Variación de Pearson de las precisiones medias de todos los motores con la operación de intersección.Fuente: elaboración propia.

En esta tabla puede observarse cómo los motores con mejores valores de *precisión* obtienen menores valores de este coeficiente, (*Google* y *MSN* en este caso), lo que indica que los valores de *precisión* obtenidos por el motor en las treinta búsquedas analizadas se encuentran más concentrados, es decir, que el rendimiento del motor es más o menos estable.

Por el contrario, aquellos motores de rendimiento discreto en cuanto a su *precisión* (*Alta Vista* y *WISEnut*), han obtenido los valores más altos en este coeficiente.

La evolución de los valores obtenidos del Coeficiente de Pearson para cada motor cuando se analizan diez, veinte o treinta documentos, presenta generalmente, un comportamiento común: existe una mayor dispersión cuando se han analizado los diez primeros documentos, atenuándose este valor cuando se han analizado treinta (aunque en nuestro estudio hemos tenido una excepción a esta regla, el motor *Terra*), tal como puede observarse en la Ilustración 6.2.

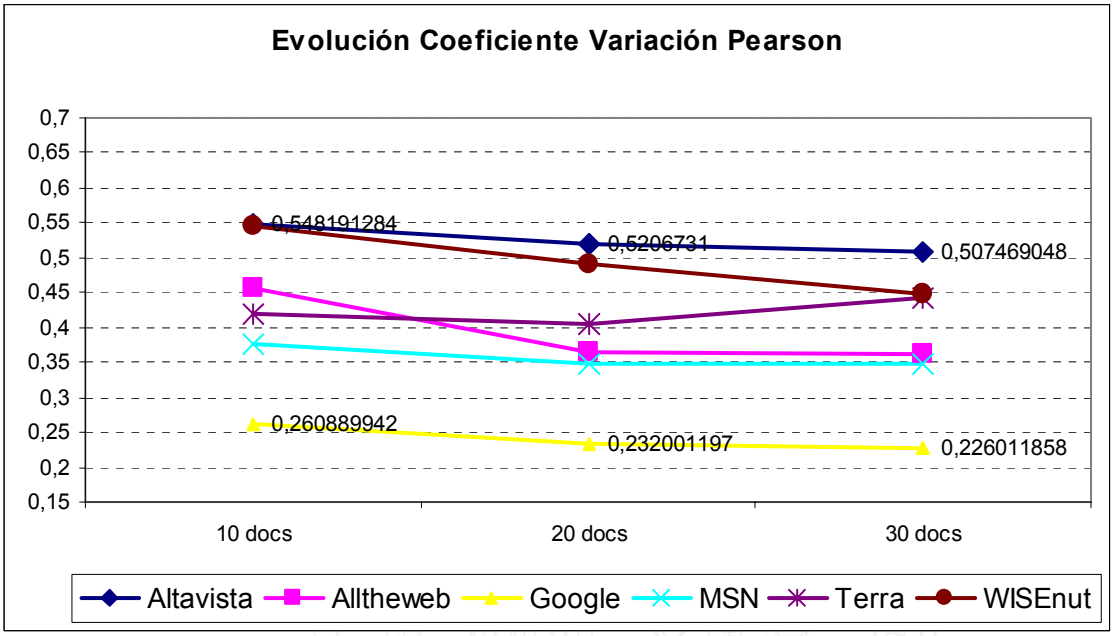


Ilustración 6.2 Evolución del Coeficiente de Variación de Pearson de las distribución de la *precisión* media de todos los motores con las operaciones combinadas de intersección. Fuente: elaboración propia.

***Precisión* de las operaciones con intersección y frase literal.**

Los resultados obtenidos con esta segunda modalidad de búsqueda han resulta muy similares a los obtenidos con la primera. Una vez más, los valores de *precisión* de *Google* son ligeramente superiores a los del resto de los motores, volviendo a ser los peores los obtenidos por el motor *Alta Vista*.

En este segundo caso, a excepción del motor *Google*, el resto se concentra en un intervalo próximo de valores, lo que implica un rendimiento más o menos similar[135].

Este conjunto de resultados (Tabla 6.5) ha sido, al igual que en el caso anterior, validado por medio de un contraste paramétrico con el fin de analizar la representatividad de los valores medios obtenidos para representar la globalidad de valores proporcionados por cada motor de búsqueda.

[135] Este comportamiento más uniforme de los motores se debe a que *Alta Vista* obtiene una serie de mejores resultados.

	10 docs	20 docs	30 docs
Altavista	0,51	0,47	0,44
Alltheweb	0,57	0,53	0,49
Google	0,75	0,69	0,64
MSN	0,62	0,55	0,51
Terra	0,57	0,53	0,49
WISEnut	0,54	0,47	0,43

Tabla 6.5 Precisiones medias de todos los motores con las operaciones combinadas de intersección y búsqueda por frase literal. Fuente: elaboración propia.

Al igual que en caso anterior, los valores de *precisión* son mayores al analizar los diez primeros documentos y descienden progresivamente al aumentar el tamaño de la búsqueda, como se ve en la Ilustración 6.3:

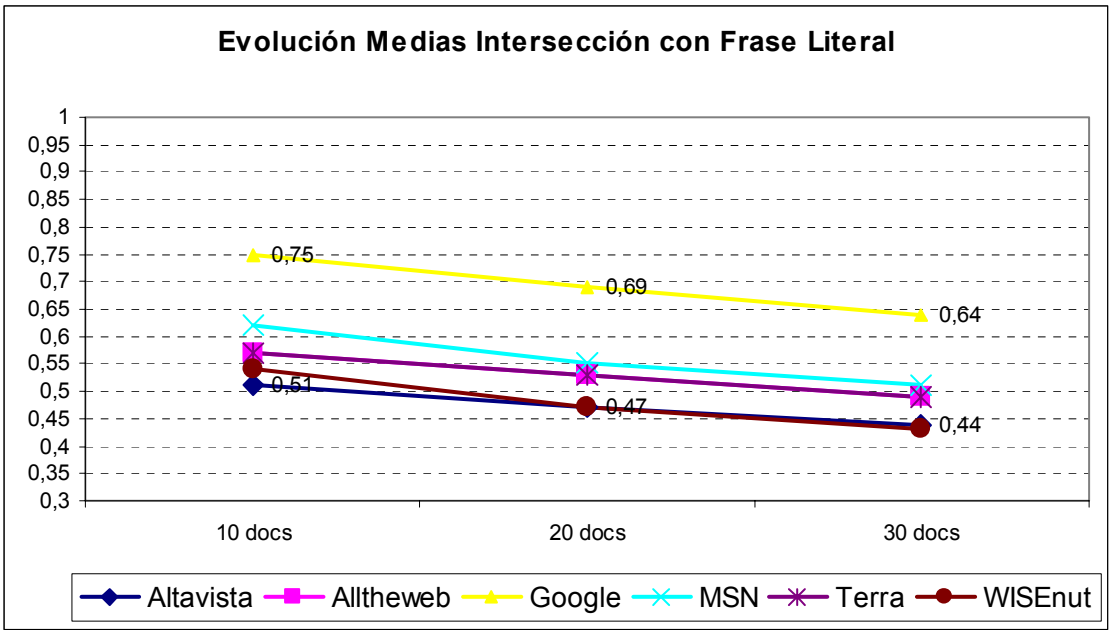


Ilustración 6.3 Evolución de las precisiones medias de todos los motores con las operaciones combinadas de intersección y búsqueda por frase literal. Fuente: elaboración propia.

El rango de valores obtenido por *Google* es ligeramente inferior al caso anterior (entre el 75% de máximo y 64% de mínimo), aunque es de similar recorrido (diez puntos de porcentaje en el primero y once en el segundo).

El cálculo del Coeficiente de Variación de Pearson para esta distribución de medias (Tabla 6.6), reitera la relación detectada en el caso anterior: mejores valores de *precisión* implican menores valores de dispersión. Es decir, un rendimiento del motor más constante.

	10 docs	20 docs	30 docs
Altavista	0,5462	0,5180	0,5105
Alltheweb	0,4257	0,3903	0,3932
Google	0,2813	0,2843	0,2955
MSN	0,4440	0,4548	0,4615
Terra	0,4966	0,4674	0,4618
WISEnut	0,4779	0,4811	0,4891

Tabla 6.6 Coeficiente de Variación de Pearson de las precisiones medias de todos los motores con las operaciones combinadas de intersección y búsqueda por frase literal. Fuente: elaboración propia.

La evolución de este porcentaje en este caso es algo más desigual que en el caso precedente, tal como refleja la Ilustración 6.4.

BIBLIOTECA VIRTUAL

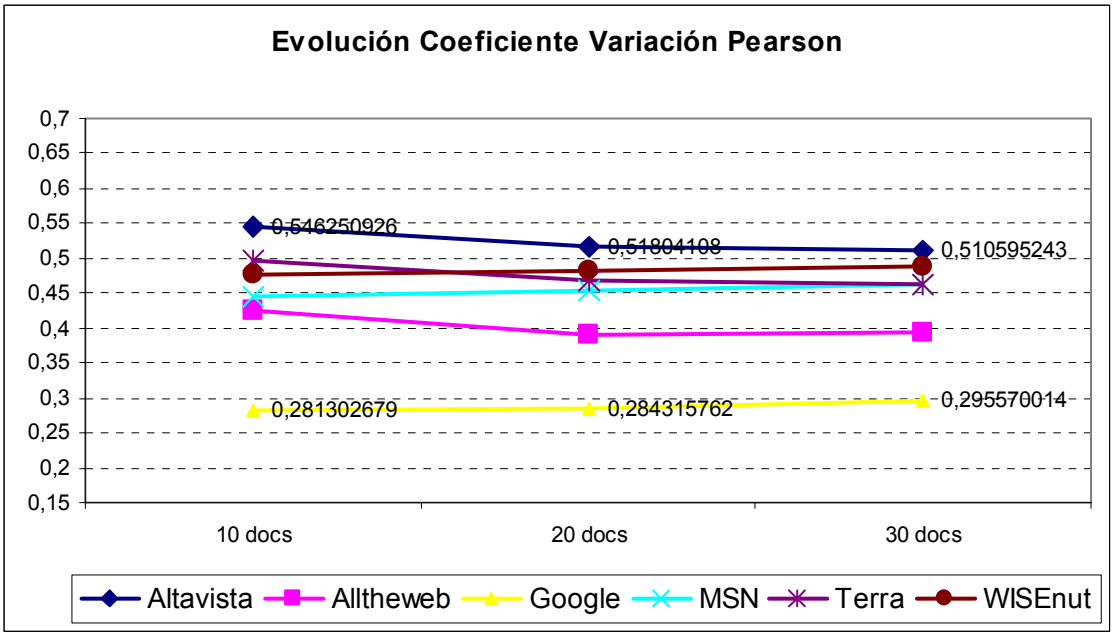


Ilustración 6.4 Evolución del Coeficiente de Variación de Pearson de las precisiones medias de todos los motores con las operaciones combinadas de intersección y búsqueda por frase literal. Fuente: elaboración propia.

Tal como puede verse, *Google* presenta unos valores casi constantes de este coeficiente en los tres análisis realizados (a los diez, veinte y treinta primeros documentos recuperados), aunque la tendencia ahora es de crecimiento ligeramente positiva, hecho que se manifiesta también en otros motores, aunque no en la totalidad.

Comparativa de las medias de los valores obtenidos.

Junto a toda la serie de datos anteriores, queda por realizar un proceso de cálculo que permita dilucidar cuál de los dos métodos de búsqueda ofrece, en términos de media aritmética de la *precisión* de los motores, unos mejores resultados.

Si a los valores obtenidos de *precisión* media para cada motor con el segundo tipo de búsqueda le restamos los valores obtenidos con el primer tipo, se obtienen los resultados que se han dispuesto en la Tabla 6.7:

	10 docs	**20 docs**	**30 docs**
Alta Vista	0,17	0,2	0,21
Alltheweb	-0,01	-0,04	-0,05
Google	-0,05	-0,05	-0,05
MSN	-0,09	-0,1	-0,09
Terra	-0,1	-0,1	-0,07
WISEnut	-0,01	-0,03	-0,03
% medio variación	-0,01	-0,01	-0,01

Tabla 6.7 Comparación de las precisiones medias de todos los motores obtenidas con ambas modalidades de búsqueda. Fuente: elaboración propia.

La Tabla 6.7 refleja que no existe prácticamente variación, en términos de *media absoluta de todos los motores*, en los niveles de *precisión* obtenidos por cada uno de ellos según las dos modalidades de búsqueda empleadas, con excepción del motor *Alta Vista* que mejora sus resultados alrededor del veinte por ciento, tal como hemos indicado anteriormente.

Datos relacionados con la *Exhaustividad*

Para calcular la *exhaustividad* relativa, según el procedimiento de Salton [SAL, 1983], es preciso disponer de un valor del tamaño de la colección, con el fin de poder verificar hasta qué grado un motor ha devuelto la mayor parte de documentos de la misma. Evidentemente, este valor no puede conocerse a priori, por lo que debe estimarse. En nuestro caso consideraremos que ese tamaño será igual a la suma de los documentos

relevantes obtenidos por los seis motores de búsqueda, eliminando los duplicados.

Así pues, el valor de *exhaustividad* que se obtenga para un motor, reflejará en gran medida, el porcentaje de documentos relevantes que ese motor está en condiciones de ofrecer como respuesta a sus usuarios en relación con la globalidad de su colección.

***Exhaustividad* de las operaciones con intersección.**

Los valores de *exhaustividad* media obtenidos en este análisis se recogen en la Tabla 6.8:

	10 docs	**20 docs**	**30 docs**
Altavista	0,09	0,08	0,09
Alltheweb	0,13	0,11	0,13
Google	0,18	0,15	0,18
MSN	0,15	0,12	0,13
Terra	0,14	0,12	0,14
WISEnut	0,12	0,09	0,11

Tabla 6.8 *Exhaustividades* medias de todos los motores con las operación de la intersección. Fuente: elaboración propia.

De similar forma que en el caso de la *precisión*, todos estos valores de *exhaustividades* medias han sido contrastados por medio de la aplicación de la prueba de T de Student.

De nuevo, el motor *Google* obtiene mejores resultados que el resto en los tres tamaños de la muestra analizada (10, 20 y 30 documentos).

Se identifica también otro comportamiento general de los seis motores analizados, ya que todos siguen una tendencia a mostrar resultados similares de *exhaustividad* cuando la muestra es de un tamaño de 10 y de 30 documentos, obteniéndose valores inferiores de esta medida en el caso de muestras de 20 documentos, tal como puede verse en la Ilustración 6.5.

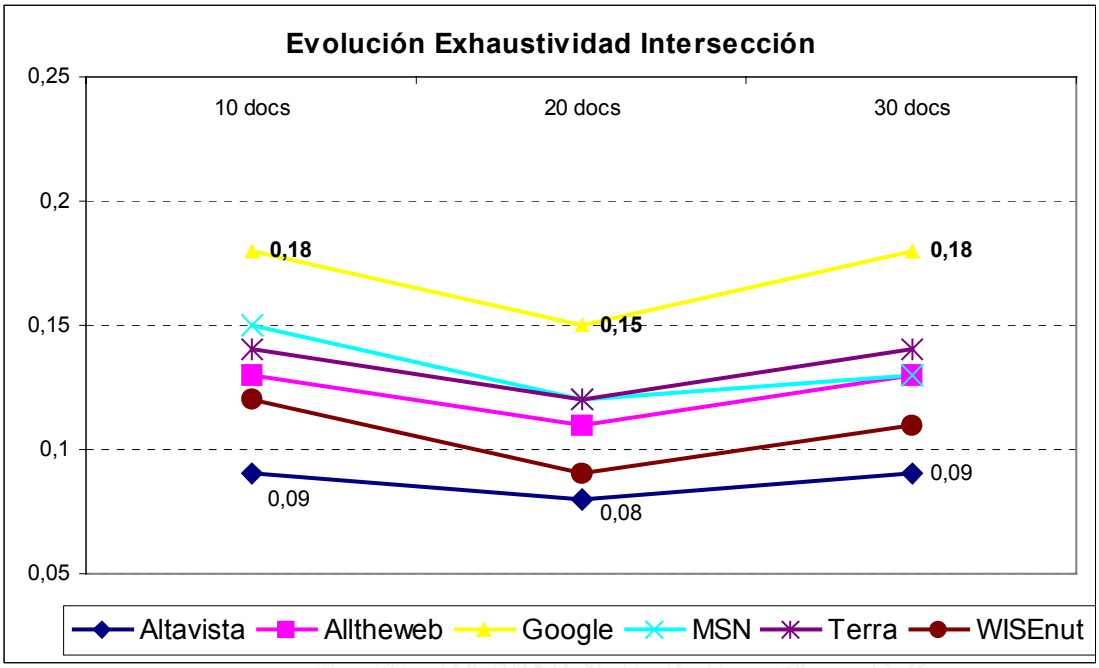


Ilustración 6.5 Evolución de las *exhaustividades* medias de todos los motores con las operación de la intersección. Fuente: elaboración propia.

Al igual que en el caso de la *precisión*, el motor *Google* vuelve a ser el que mejores resultados presenta y el motor *Alta Vista* reitera su mal comportamiento, situación lógica en tanto que ambas medidas (*exhaustividad* y *precisión*) dependen del número de documentos relevantes obtenidos por cada motor.

El comportamiento del resto de los motores es más o menos similar, agrupando sus resultados en una zona intermedia entre los dos motores citados.

Estos resultados de *exhaustividad* obtenidos por los distintos motores, indican que será el motor *Google* el que obtenga los mejores resultados cuando se aplique posteriormente la medida de valor I_1 simple de Borko, tal como se verá posteriormente, ya que en el cálculo de la *precisión* también obtuvo los mejores resultados.

Si se procede a continuación a determinar el Coeficiente de Variación de Pearson de la distribución de las *exhaustividades* medias de cada motor de búsqueda, se observará que los resultados obtenidos por el motor *Google* lo convierten en el motor de menor dispersión (es decir, el motor que presenta un comportamiento más estable), aunque, a medida que ha ido aumentando el número de elementos de la muestra hasta los treinta documentos, todos los motores (excepto *Alta Vista*) se han ido agrupando

en torno a un mismo valor (tal como puede observarse en la Ilustración 6.6), evolución que no se presentaba en el caso de la medida de la *precisión*.

La circunstancia que sí se ha repetido, con respecto a este coeficiente y la operación de *precisión*, es que las distribuciones se encuentran más dispersas cuando la muestra de los documentos analizados es mayor.

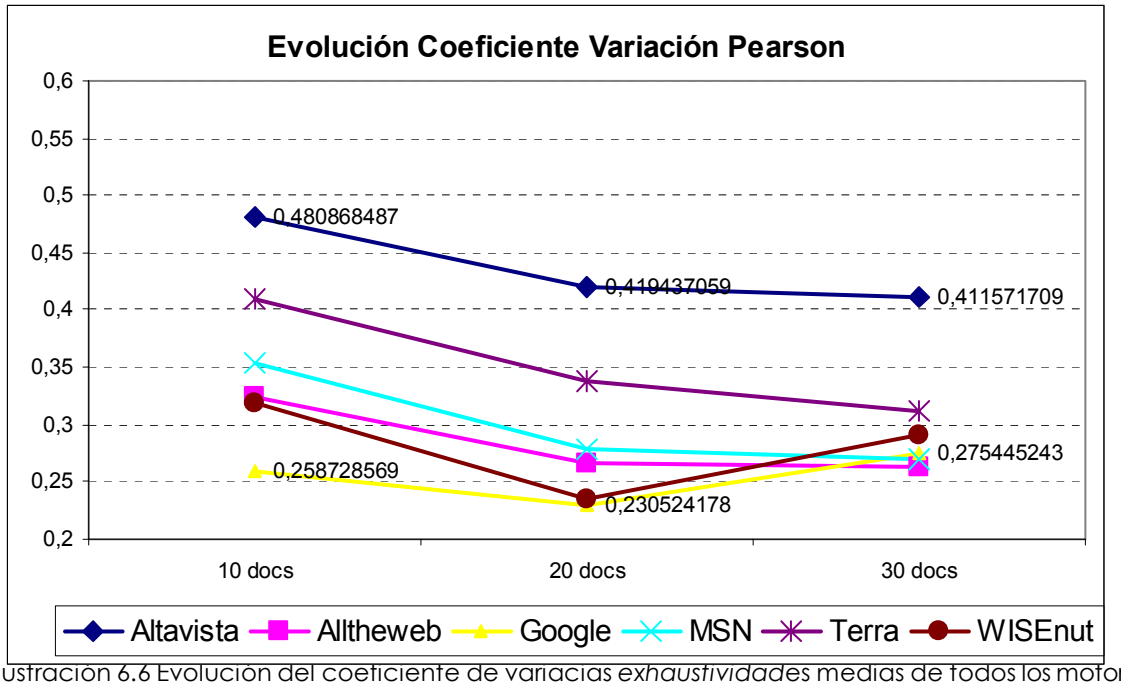


Ilustración 6.6 Evolución del coeficiente de variacias *exhaustividades* medias de todos los motores con las operación de la intersección. Fuente: elaboración propia.

***Exhaustividad* de las operaciones con intersección y frase literal.**

Los valores de *exhaustividad* media obtenidos con esta segunda modalidad de búsqueda se recogen en la Tabla 6.9:

	10 docs	**20 docs**	**30 docs**
Altavista	0,12	0,11	0,13
Alltheweb	0,14	0,12	0,15
Google	0,19	0,17	0,2
MSN	0,14	0,12	0,13
Terra	0,13	0,12	0,14
WISEnut	0,12	0,09	0,11

Tabla 6.9 *Exhaustividades* medias de todos los motores con las operaciones combinadas de intersección y búsqueda por frase literal. Fuente: elaboración propia.

La Tabla 6.9 muestra que el motor *Alta Vista* mejora ligeramente sus resultados con respecto al primer tipo de operación, aunque el resto de los motores presenta unos niveles de *exhaustividad*es medias similares.

En la Tabla 6.10 se muestran las diferencias que se establecen entre los dos tipos de búsqueda para todos los motores (al igual que con las precisiones se va a restar el valor del primer tipo de búsqueda al resultado obtenido con el segundo tipo).

	10 docs	**20 docs**	**30 docs**	**Medias**
Altavista	0,03	0,03	0,04	0,03
Alltheweb	0,01	0,01	0,02	0,01
Google	0,01	0,02	0,02	0,01
MSN	-0,01	0	0	0
Terra	-0,01	0	0	0
WISEnut	0	0	0	0
			Media global	0,009

Tabla 6.10 Diferencia de las *Exhaustividad*es medias obtenidas entre los dos tipos de búsqueda. Fuente: elaboración propia.

Los resultados ofrecidos por la Tabla 6.10 son concluyentes, para el parámetro de la *exhaustividad* tampoco existe prácticamente ninguna diferencia entre un tipo y otro de búsqueda.

Es por ello que no se ha considerado necesario calcular la evolución del Coeficiente de Variación de Pearson para esta distribución.

Datos sobre total de documentos localizados.

El número total de documentos recuperados por los motores operando con la intersección es muy superior (como no podía ser de otro modo), al número de documentos recuperados con la segunda modalidad de búsqueda (que menos exhaustiva).

La Tabla 6.11 muestra claramente estas diferencias y las medias de cada modalidad de búsqueda contrastadas por medio de la prueba "T" de Student (ver Anexo):

	INT	INT + FRA	PORC	MED (INT)	MED(INT+FRA)
Altavista	85393	10163	840,23	2846	339
Alltheweb	157143	18926	830,30	5238	631
Google	252191	38298	658,49	8248	1277
MSN	87842	10085	871,01	2928	336
Terra	81053	9032	897,39	2702	301
WISEnut	77481	8128	953,26	2583	271

Tabla 6.11 Total de documentos localizados por cada motor en las treinta preguntas realizadas con cada modalidad de búsqueda. INT: operación intersección. INT + FRA: operación combinada de intersección y frase literal. PORC: INT dividido por INT + FRA y multiplicado por cien. MED(INT): media de documentos devueltos en operación intersección. MED (INT+FRA): media de documentos devueltos en operación combinada intersección y frase literal. Fuente: elaboración propia.

Esta tendencia no se repite, en cambio, si nos limitamos únicamente a calcular el total de documentos relevantes distintos calculados por los seis motores en cada caso (Tabla 6.12).

En este caso, las cifras se asemejan mucho más (1911 documentos frente a 1796), lo que induce a pensar que, independientemente de la modalidad de búsqueda empleada, los algoritmos de alineamiento de los motores de búsqueda tienden a mostrar entre los primeros treinta documentos devueltos, una serie de documentos relevantes de dimensiones similares.

Búsqueda	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Subtotal I**
INT	115	41	108	80	78	45	80	67	68	83	765
INT + FRA	87	71	84	115	70	39	71	62	75	64	738
Búsqueda	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**Subtotal II**
INT	43	56	111	34	80	74	66	67	47	78	1421
INT + FRA	54	103	111	13	89	74	53	46	42	63	1386
Búsqueda	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**Total**
INT	45	48	42	42	59	57	72	55	32	38	1911
INT + FRA	32	10	8	58	54	61	62	59	28	38	1796

Tabla 6.12 Total de documentos relevantes distintos localizados por cada motor en las treinta preguntas realizadas con cada modalidad de búsqueda. INT: operación intersección. INT + FRA: operación combinada de intersección y frase literal. Fuente: elaboración propia.

Datos sobre enlaces erróneos y duplicados.

Paralelamente a la toma de datos sobre la *relevancia* de las páginas de las respuestas de cada motor de búsqueda, se ha anotado todo el conjunto de incidencias relacionadas con la detección de un enlace erróneo o duplicado, siguiendo para ello las indicaciones recogidas en la propuesta de evaluación, sin apartarse de las mismas en ningún momento, con el fin de no introducir distorsiones en el experimento.

Datos sobre enlaces erróneos.

En este análisis sólo se ha anotado la presencia de un tipo de enlace erróneo dentro de la respuesta de cada motor. No se anotado la causa de este error y tampoco se han señalado diferencias entre los posibles tipos de errores, como sí hace Notess quien cuenta por separado los errores totales y los errores debidos a fallos "tipo error 400" del protocolo HTTP[136] [NOT, 2000c]. Los datos obtenidos sobre este parámetro, se recogen en la Tabla 6.13.

	INT	% INT	INT + FRA	% INT + FRA	DIF
Altavista	2,54	8,46	2,41	8,03	-0,43
Alltheweb	2,2	7,33	2,2	7,33	0
Google	1,06	3,53	1,44	4,8	1,27
MSN	1,65	5,5	1,1	3,66	-1,84
Terra	2,63	8,76	2,26	7,53	-1,23
WISEnut	5,56	18,53	4,13	13,76	-4,77

Tabla 6.13 Porcentajes de enlaces erróneos proporcionado por cada motor en el proceso de evaluación tras la realización de las treinta búsquedas. INT: resultados con operación intersección. %INT: porcentaje con operación intersección. INT + FRA: resultados con operación combinada intersección y frase literal. %INT + FRA: porcentaje con operación combinada intersección y frase literal DIF: diferencia entre el segundo y el primer porcentaje. Fuente: elaboración propia.

Los datos recogidos en las columnas primera y tercera de esta tabla, corresponden al número medio de errores sobre las treinta direcciones que conformaban las muestras analizadas. Estos valores se han determinado en

[136] Este error se produce cuando la página no se encuentra en el servidor web bien tiene su acceso restringido. El protocolo HTTP diferencia otros mensajes de error, más información en: <http://www.um.es/gtiweb/fjmm/ttsite/tema4-parte2.htm#http-codigos>

tantos por ciento medios y se recogen en la segunda y cuarta columna. La quinta columna determina la diferencia de porcentaje de error entre un tipo de búsqueda y otro (es el resultado de restar al porcentaje de fallos de la segunda modalidad de búsqueda el porcentaje obtenido por la primera).

De este análisis pueden extraerse los siguientes resultados:

- Sólo *Google* y Microsoft Network obtienen porcentajes por debajo o próximos al 5% en ambas modalidades de búsqueda.
- El resto se agrupa en un intervalo comprendido entre el 7% y 9%, con excepción de *WISEnut* que obtiene valores muy altos.
- Este comportamiento se repite en ambas modalidades de búsqueda.
- De hecho no existen diferencias significativas en los resultados obtenidos por cada motor en ambas modalidades.

Tal como se desprende de los porcentajes recogidos en la tabla 5.13 y se muestra gráficamente en la Ilustración 6.7, los resultados del motor *WISEnut* son muy negativos, lo que implica necesariamente peores valores en el análisis de la *precisión*. Llama poderosamente la atención que el índice de este motor parece no actualizarse, situación harto sorprendente al tratarse de un motor de reciente implantación y basado en la extensión de enlaces.

Al hecho probado de la gran cantidad de errores de refresco de su índice hay que sumar el que, desde hace varios meses[137], su página principal anuncia que dispone de 1.571.413.207[138] páginas en su índice, lo cual induce a pensar en esa ausencia de actualización de este índice.

[137] Período comprendido entre septiembre de 2001 y abril de 2002.

[138] En el capítulo dedicado a la evaluación de los SRI en la web se ha visto que Notess discrepa sustancialmente de esta cifra, rebajándola a algo más de 500 millones de páginas [NOT, 2000a].

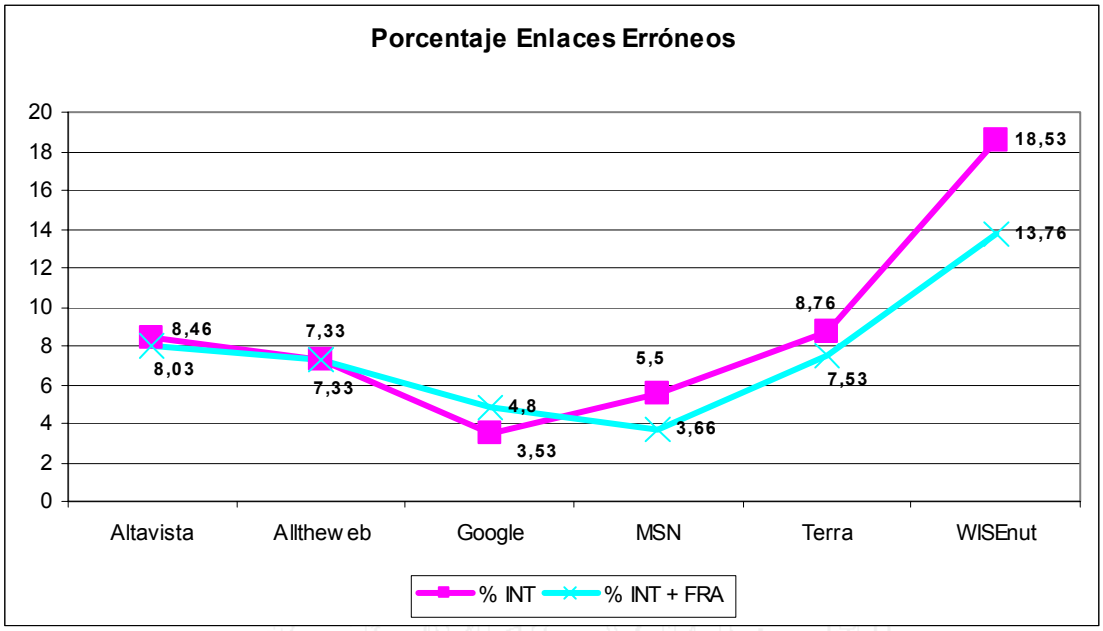


Ilustración 6.7 Evolución de los porcentajes de enlaces erróneos proporcionados por cada motor en ambas modalidades de búsqueda. Fuente: elaboración propia.

El citado estudio realizado por Notess [NOT, 2000c], incluye cuatro de los motores analizados en nuestro experimento, y los resultados obtenidos presentan coincidencias, tal como puede observarse en la Tabla 6.14:

	% Error	% Error Notess	% Error Notess (tipo 400)
Altavista	8,03	13,7	9,3
Alltheweb	7,33	2,3	1,8
Google	3,53	4,3	3,3
MSN	3,66	1,7	1,0

Tabla 6.14 Comparación porcentajes enlaces erróneos en relación al estudio de Notess. Fuente: elaboración propia, Notess, G. Search Engine Statistics: Dead reports. [En línea]. Bozeman, MT: Notes.com, 2000. <http://www.searchengineshowdown.com/stats/dead.shtml> [Consulta: 12 marzo 2002]

Para la realización de esta tabla se ha elegido el porcentaje más favorable para cada motor en las dos modalidades de búsqueda analizadas.

En líneas generales, los datos del estudio de Notess son más benévolos con los motores, siendo sus porcentajes de error algo menores, con la excepción del motor *Google* que obtiene resultados casi similares.

El motor *Alta Vista* vuelve a ser el que mayor porcentaje de fallos obtiene. No obstante, hay que tener en cuenta que el estudio de Notess es de febrero de 2000 y los índices de los motores han crecido sustancialmente desde esa fecha y por lo tanto, resulta lógica esta variación en los porcentajes, habida cuenta de la ya comentada naturaleza dinámica de la web.

Datos sobre enlaces duplicados.

La propuesta de evaluación incluye una serie de casos en los que se considera como duplicado alguna dirección devuelta por un motor de búsqueda, destacando de la misma los redireccionamientos o distintas direcciones dinámicas procedentes de consultas a un mismo portal de Internet. En este caso, no se dispone de estudios que permitan cotejar los resultados obtenidos, aunque si existieran, esa comparación siempre sería relativa porque los supuestos establecidos de forma previa al análisis de las direcciones devueltas no tienen por qué coincidir.

	INT	% INT	INT + FRA	% INT + FRA	DIF
Altavista	0,93	3,1	0,74	2,46	-0,64
Alltheweb	0,6	2	0,5	1,66	-0,34
Google	1,13	3,76	0,68	2,26	-1,5
MSN	0,79	2,63	0,73	2,43	-0,2
Terra	1,66	5,53	1,7	5,66	0,13
WISEnut	0,26	0,86	0,33	1,1	0,24

Tabla 6.15 Porcentajes de enlaces duplicados proporcionados por cada motor en el proceso de evaluación. INT: resultados con operación intersección. %INT: porcentaje con operación intersección. INT + FRA: resultados con operación combinada intersección y frase literal. %INT + FRA: porcentaje con operación combinada intersección y frase literal DIF: diferencia entre el segundo y el primer porcentaje.Fuente: elaboración propia.

La distribución de las columnas en la Tabla 6.15 es similar a la realizada en el caso anterior de los enlace duplicados, volviendo a observarse escasas diferencias entre una y otra modalidad de búsqueda.

Casi todos los motores establecen sus porcentajes entre el uno y el tres por ciento, siendo la excepción el motor *Terra* que fija sus valores entre el cinco y el seis por ciento (Ilustración 6.8).

En todo caso, estos porcentajes podrían haber variado si en el desarrollo del proceso de análisis se hubieran retocado los criterios establecidos de

valoración, en tanto que surgieron dos circunstancias que podrían haber provocado esta modificación.

La primera de estas circunstancias es la tendencia, que tienen algunos motores, a devolver como resultado secciones del mismo directorio correspondientes a versiones de diferentes países, por ejemplo, las direcciones **<http://es.dir.yahoo.com/ciencia_y_tecnologia/ingenieria/>** y **<http://ar.dir.yahoo.com/ciencia_y_tecnologia/ingenieria/>**; que hacen referencia en la práctica al mismo recurso, el directorio de *Yahoo*, aunque cada dirección es diferente e incluso la interface de usuario es distinta en cada caso.

Aplicando los criterios establecidos en la propuesta de evaluación, ambas direcciones deberían ser reducidas a su forma más simple, es decir, **<http://es.dir.yahoo.com>** y **<http://ar.dir.yahoo.com>,** y como ninguna de estas dos direcciones puede ser considerada como redireccionamiento de la otra, deben considerarse como dos direcciones distintas y relevantes (siempre y cuando en el directorio aparezcan referencias a sitios web relacionados con la materia objeto de la temática, que en el caso de *Yahoo* suele ser normal).

Si de alguna manera se estipulara que se trata de un enlace duplicado o, dicho de forma más precisa, un enlace "redundante" (que realmente no sería un duplicado aunque se reincida el envío a una sección del mismo directorio, aunque la determinación de cuál es la original y cuál es la redundante no sería tarea simple), y alguna de las dos direcciones fuera considerada de *relevancia* igual a cero o bien se obviara su presencia en la lista de respuesta y se procediera a analizar otro recurso, los valores E-P obtenidos por ese motor serían, por regla general, algo menores que los obtenidos en este estudio y aumentarían los porcentajes de enlaces duplicados en la misma proporción.

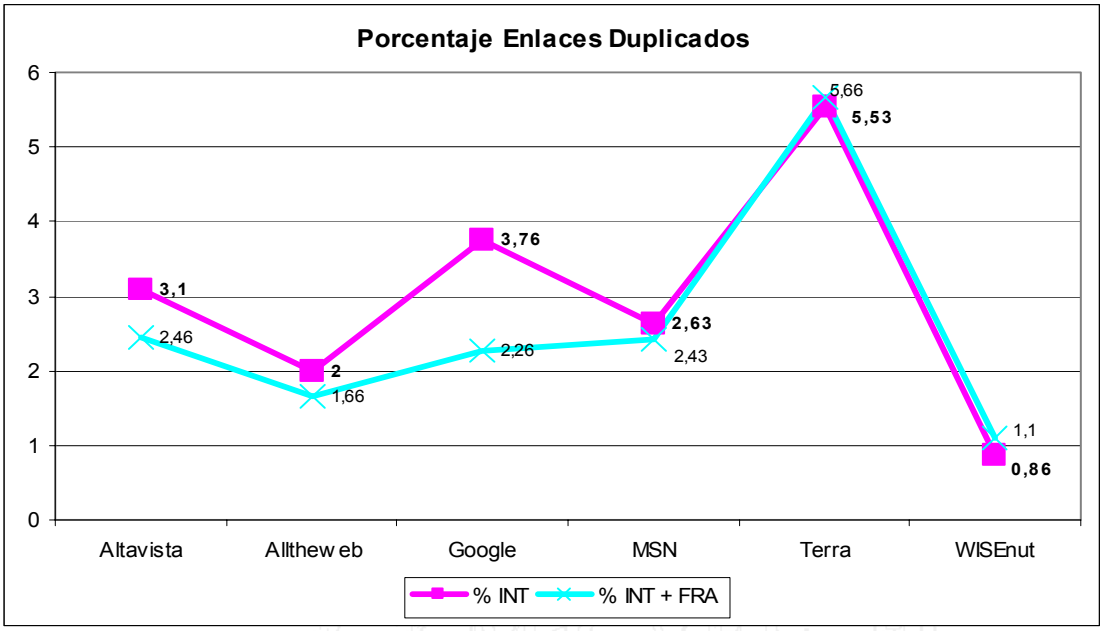


Ilustración 6.8 Evolución de los porcentajes de enlaces duplicados proporcionados por cada motor en ambas modalidades de búsqueda. Fuente: elaboración propia.

La segunda circunstancia que surgió fue la verificación, a lo largo del proceso de análisis, de que la interface de usuario de cada motor muestra de manera diferente los resultados.

Algunos motores muestran una dirección de un sitio web (por ejemplo **<http://www.um.es/pjumilla/>**) e indican al usuario si quiere consultar más direcciones del mismo sitio web, es decir, agrupan alrededor de una dirección "pivote" todos los recursos de los que disponen de ese sitio web. Otros motores, en cambio, si además de esta dirección "pivote" poseen alguna más, las incluyen en la respuesta como recursos independientes.

Esta circunstancia tampoco se encontraba prevista en la propuesta de evaluación, por lo que no se pudo tomar ninguna acción al respecto aunque, al igual que en el primer caso, si de alguna manera se estipulara que un enlace fuera "redundante" y algunas de las direcciones fueran consideradas de *relevancia* igual a cero o bien se obviara su presencia en la respuesta del motor, procediendo a analizar otro recurso, los pares de valores E-P obtenidos por ese motor serían menores que los obtenidos en este estudio y aumentarían los porcentajes de enlaces duplicados en la misma proporción. Tanto en una como en otra circunstancia, si se introdujeran criterios destinados a paliar su efecto, los motores *Microsoft Netrwork*, *Terra* y *WISEnut* se verían muy perjudicados con respecto a los otros tres, que no padecen tanto este problema.

Datos sobre la similitud de los resultados.

El análisis de la similitud de los resultados obtenidos en nuestro experimento se ha realizado desde diversas perspectivas. En primer lugar se han calculado las medias de estas similitudes obtenidas al analizar respectivamente los diez, veinte y treinta primeros documentos, con el fin de calcular en cuál de los tres tamaños de la muestra es mayor la coincidencia de resultados (en principio, cabría esperar que con un mayor número de documentos analizados, el valor de la similitud sería mayor).

En segundo lugar se han comparado de forma individual las similitudes obtenidas por cada motor en los treinta casos analizados, con el fin de verificar si su comportamiento (en lo relacionado con la similitud) sigue un patrón establecido o no.

Por último, en tercer lugar, se han comparado los conjuntos de documentos relevantes recuperados por ambas modalidades de búsqueda para los treinta temas analizados, con el fin de precisar si estos conjuntos, además de ser de dimensiones parecidas, son de similares contenidos.

Medias de las similitudes obtenidas.

La similitud calculada no sólo mide la coincidencia de resultados ofrecidos por los distintos motores, tal como calculó Lojslund [LOJ, 2000], sino que, por la aplicación de la función de ponderación de la *relevancia* que se ha definido, también sirve para reflejar la semejanza existente en el orden de presentación de los documentos.

Los resultados obtenidos en este cálculo verifican la hipótesis planteada originalmente: se obtiene un mayor grado de coincidencia a medida que aumenta el tamaño de la muestra analizada, tal como se muestra en la Tabla 6.16:

	SIM(INT)	% AUMENTO	SIM(INT + FRA)	% AUMENTO	DIF
10 docs	0,134		0,151		0,017
20 docs	0,145	8,2	0,17	12,6	0,025
30 docs	0,157	8,7	0,184	8,7	0,027

Tabla 6.16 Similitudes medias obtenidas en el experimento. SIM(INT): Similitud media en las operaciones de intersección. %AUMENTO: Porcentaje de crecimiento de la similitud media. SIM(INT + FRA): Similitud media en las operaciones combinadas de intersección y frase literal. DIF: diferencia entre las similitudes medias del segundo y el primer tipo de búsqueda. Fuente: elaboración propia.

La Tabla 6.16 muestra un crecimiento en torno al 8% de los valores de similitud media, a medida que crece la muestra de documentos analizada.

Esta tabla también refleja que los valores medios de similitud existentes entre los seis motores de búsqueda analizados alcanzan cotas inferiores al 20%, lo que explica la amplia divergencia de resultados que suelen ofrecer. En último lugar, se observa claramente que la modalidad de búsqueda empleada no afecta a los valores de similitud que permanecen casi constantes (tal como se aprecia en la última columna de la Tabla 6.16).

Medias de las similitudes obtenidas motor a motor.

En la Tabla 6.17 se han recogido los valores medios de las similitudes obtenidas para cada par de motores con muestras de diez, veinte y treinta documentos analizados y la operación de intersección:

Docs	AW-GO	AW-MS	AW-TE	AW-WI	AV-AW	AV-GO	AV-MS	AV-TE	AV-WI	GO-MS	GO-TE	GO-WI	MS-TE	MS-WI	TE-WI
10	0,21	0,10	0,12	0,16	0,13	0,10	0,05	0,05	0,12	0,15	0,17	0,15	0,30	0,13	0,15
20	0,22	0,12	0,14	0,16	0,14	0,12	0,06	0,06	0,11	0,17	0,17	0,18	0,30	0,15	0,16
30	0,23	0,12	0,15	0,17	0,14	0,13	0,07	0,07	0,12	0,19	0,20	0,20	0,30	0,17	0,16

Tabla 6.17 Similitudes medias obtenidas en el experimento por cada par de motores, con 10, 20 y 30 documentos analizados con la operación de la intersección. AW: *All the Web*; AV: *Alta Vista*; GO: *Google*; MS: Microsoft Network; TE: *Terra*; WI: *WISEnut*. Fuente: elaboración propia.

Como puede apreciarse, los valores de las similitudes apenas varían al modificar el tamaño de la muestra, manteniéndose en unas dimensiones muy constantes.

Los motores más similares (por lo tanto, más próximos en un espacio vectorial), son los motores *MSN* y *Terra* (circunstancia lógica ya que comparten un porcentaje elevado de la colección y emplean el mismo motor: *Inktomi*).

Estos dos motores, al mismo tiempo, presentan la misma distancia con el motor *Alta Vista*, siendo este valor el mínimo de toda la distribución, es decir, se encuentran igual de alejados de *Alta Vista* y esta es la distancia más grande que se ha detectado.

La Tabla 6.18 recoge la distancia media de cada motor con respecto a todos los demás, para su cálculo se ha obtenido la media obtenidas por un motor con respecto a los otros cinco (con los valores correspondientes a la muestra de treinta documentos).

	SIM	DIST
Altavista	0,142	0,858
Alltheweb	0,106	0,894
Google	0,19	0,81
MSN	0,17	0,83
Terra	0,176	0,824
WISEnut	0,172	0,828

Tabla 6.18 Similitud y Distancia media de cada motor con respecto al resto; operación intersección. SIM: Similitud media. DIST: distancia media (1 – SIM). Fuente: elaboración propia.

En el caso de las búsquedas basadas en la combinación de la operación de intersección y de frase literal, sus resultados se recogen en la Tabla 6.19, una vez calculados para cada par de motores y con muestras de diez, veinte y treinta documentos analizados.

Docs	AW-GO	AW-MS	AW-TE	AW-WI	AV-AW	AV-GO	AV-MS	AV-TE	AV-WI	GO-MS	GO-TE	GO-WI	MS-TE	MS-WI	TE-WI
10	0,12	0,15	0,13	0,15	0,14	0,12	0,13	0,07	0,18	0,18	0,15	0,17	0,24	0,19	0,12
20	0,21	0,15	0,16	0,16	0,16	0,15	0,15	0,11	0,19	0,19	0,17	0,20	0,29	0,19	0,13
30	0,23	0,16	0,17	0,18	0,19	0,17	0,16	0,13	0,12	0,19	0,19	0,21	0,36	0,23	0,19

Tabla 6.19 Similitudes medias obtenidas en el experimento por cada par de motores, con 10, 20 y 30 documentos analizados, con la operación combinada intersección y frase literal. AW: *All the Web*; AV: *Alta Vista*; GO: *Google*; MS: Microsoft Network; TE: *Terra*; WI: *WISEnut*. Fuente: elaboración propia.

La distribución de valores de las similitudes calculadas motor a motor resulta algo más variada, en algunos casos, con respecto a los resultados obtenidos con la operación de la intersección.

Los motores más similares vuelven a ser *MSN* y *Terra* (con un 0,36 de similitud frente a un 0,30 anterior).

Las similitudes mínimas muestran valores más altos (cercanos al doble) que en la operación de la intersección (0,13 *Alta Vista* con *Terra* y 0,12 *Alta Vista* con *WISEnut*), cuando las muestras son de treinta documentos, aunque entre *Alta Vista* y *Terra* se presenta la misma distancia al analizar la muestra de diez documentos.

La distribución de las similitudes y distancias medias de cada motor con respecto al resto, se recoge en la Tabla 6.20:

	SIM	DIST
Altavista	0,186	0,814
Alltheweb	0,128	0,872
Google	0,198	0,802
MSN	0,22	0,78
Terra	0,208	0,792
WISEnut	0,186	0,814

Tabla 6.20 Similitud y Distancia media de cada motor con respecto al resto; operación combinada de intersección y frase literal. SIM: Similitud media. DIST: distancia media (1 – SIM). Fuente: elaboración propia.

Si se agruparan las distancias obtenidas con ambas modalidades de búsqueda, en un gráfico radial establecido en torno a un hipotético centro de gravedad (que podría considerarse) como el núcleo de la colección de documentos de la web, se obtendría la siguiente distribución (Ilustración 6.9):

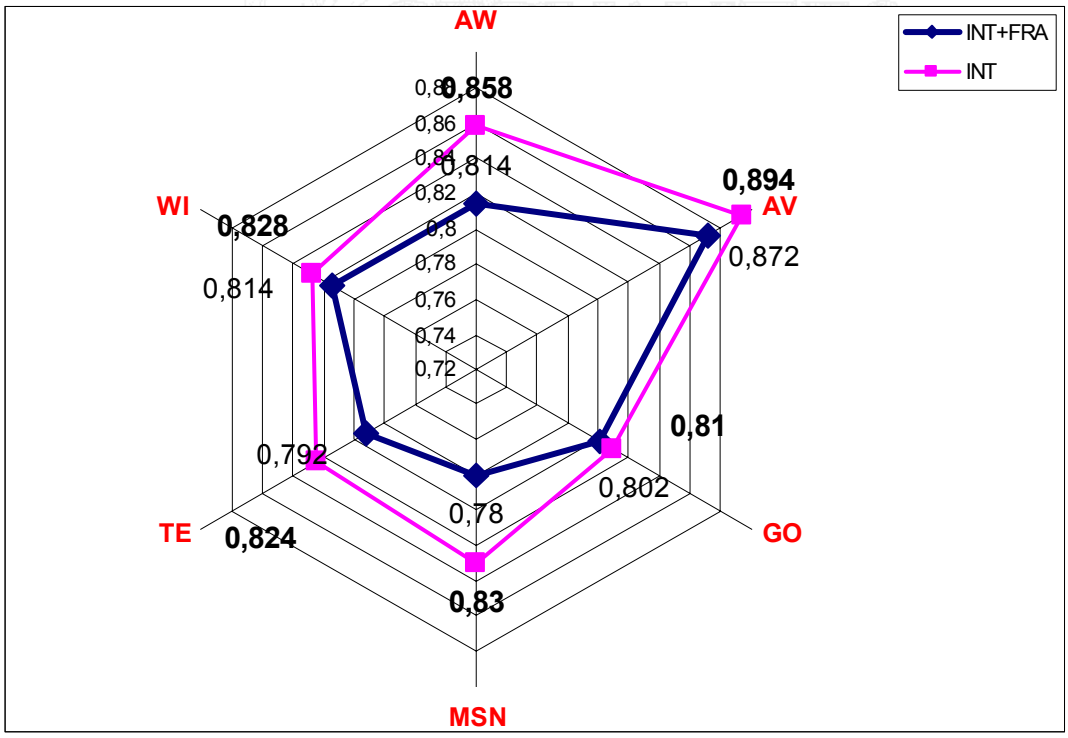


Ilustración 6.9 Distribución radial de la Distancia media de los motores de búsqueda con ambas posibilidades de búsqueda. Fuente: elaboración propia.

Si se considera, siempre de forma hipotética, que el centro del gráfico constituye la mejor colección de documentos posible de recuperar en la

web sobre los treinta temas analizados, los motores de búsqueda se encuentran más cerca de alcanzar ese centro en el caso de la segunda modalidad de búsqueda, aunque las distancias obtenidas por los motores de búsqueda evaluados son ciertamente considerables con respecto a esa colección ideal. Con la operación de intersección, el motor *Google* es el más cercano a la misma, seguido de cerca por *Terra* y *MSN*. Con la segunda posibilidad de búsqueda, el motor *MSN* y el motor *Terra* (por este orden), superan a *Google*. En ambos casos, los resultados más distantes los obtiene *Alta Vista*.

Similitud de resultados obtenidos por los motores con ambas modalidades de búsqueda en el mismo tema.

La última serie de cálculos de similitud realizada en este experimento, sirven para verificar el grado de semejanza de los resultados que ofrecen los seis motores en cada uno de los treinta temas analizados, según la modalidad de búsqueda empleada. Los resultados se recogen en la Tabla 6.21:

Tema	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		
INT	1	8	2	4	6	11	13	15	17	19		
INT + FRA	7	9	3	5	10	12	14	16	18	20		
											Σsim	**Media**
similitud	0,357	0,524	0,449	0,361	0,619	0,724	0,613	0,732	0,419	0,236	5,033	0,503
Tema	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**		
INT	21	23	25	27	29	31	33	35	37	39		
INT + FRA	22	24	26	28	30	32	34	36	38	40		
											Σsim	**Media**
similitud	0,419	0,561	0,694	0,531	0,242	0,423	0,716	0,296	0,611	0,479	10	0,5
Tema	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**		
INT	41	43	45	47	49	51	53	55	57	59		
INT + FRA	42	44	46	48	50	52	54	56	58	60		
											Σsim	**Media**
similitud	0,662	0,182	0,206	0,673	0,662	0,574	0,35	0,579	0,508	0,724	15,12	0,504

Tabla 6.21 Similitudes de los vectores de resultado obtenidos tema por tema en ambas modalidades de búsqueda. INT: búsquedas basadas en la operación de la intersección. INT + FRA: búsquedas basadas en las operaciones combinadas de intersección y frase literal. Σsim: suma parcial de las similitudes de cada fila Media: subtotal y total de medias de las similitudes. Fuente: elaboración propia.

La Tabla 6.21 muestra que las colecciones de documentos obtenidas por los seis motores en ambas modalidades de búsqueda poseen una similitud que oscila alrededor del cincuenta por ciento (0.504 más concretamente).

Anteriormente se había calculado que las dimensiones de ambas colecciones de documentos eran de tamaño similar.

Estos valores permiten deducir que la mitad de los documentos relevantes mostrados en su conjunto, por los seis motores analizados, suele ser similar, independientemente del tipo de búsqueda empleado. Al mismo tiempo, la otra mitad de documentos relevantes es diferente.

Datos sobre la medida de valor simple de Borko.

A partir de los pares de valores E-P determinados en el experimento para ambas modalidades de búsqueda (presentados en las tablas 5.3, 5.5, 5.8 y 5.9), se construye la Tabla 6.22.

Esta tabla recoge los valores medios de la medida de valor simple I_1 de Borko de cada motor en ambas modalidades de búsqueda: las basadas en la operación de intersección y las basadas en la combinación de la intersección y de las búsquedas por frase literal.

Esta medida se calcula sumando los valores obtenidos de *exhaustividad* y de *precisión* para cada uno de los casos analizados.

	I1 de Borko					
	INT			**INT + FRA**		
	10 docs	20 docs	30 docs	10 docs	20 docs	30 docs
Alta Vista	0,52	0,47	0,45	0,63	0,58	0,57
Alltheweb	0,71	0,65	0,64	0,71	0,65	0,64
Google	0,96	0,87	0,86	0,94	0,86	0,84
MSN	0,83	0,73	0,69	0,76	0,67	0,64
Terra	0,77	0,71	0,67	0,70	0,65	0,63
WISEnut	0,66	0,58	0,56	0,66	0,56	0,54

Tabla 6.22 Valores de la medida I1 de Borko para ambas modalidades de búsqueda. INT: búsquedas basadas en la operación de la intersección. INT + FRA: búsquedas basadas en las operaciones combinadas de intersección y frase literal. Fuente: elaboración propia.

Tal como se adelantaba en la presentación de las tablas de pares de valores E-P, el motor *Google* presenta los mejores valores de esta medida de valor simple en ambas modalidades de búsqueda.

La Ilustración 6.10 muestra la evolución seguida por esta medida en el caso de la operación de la intersección.

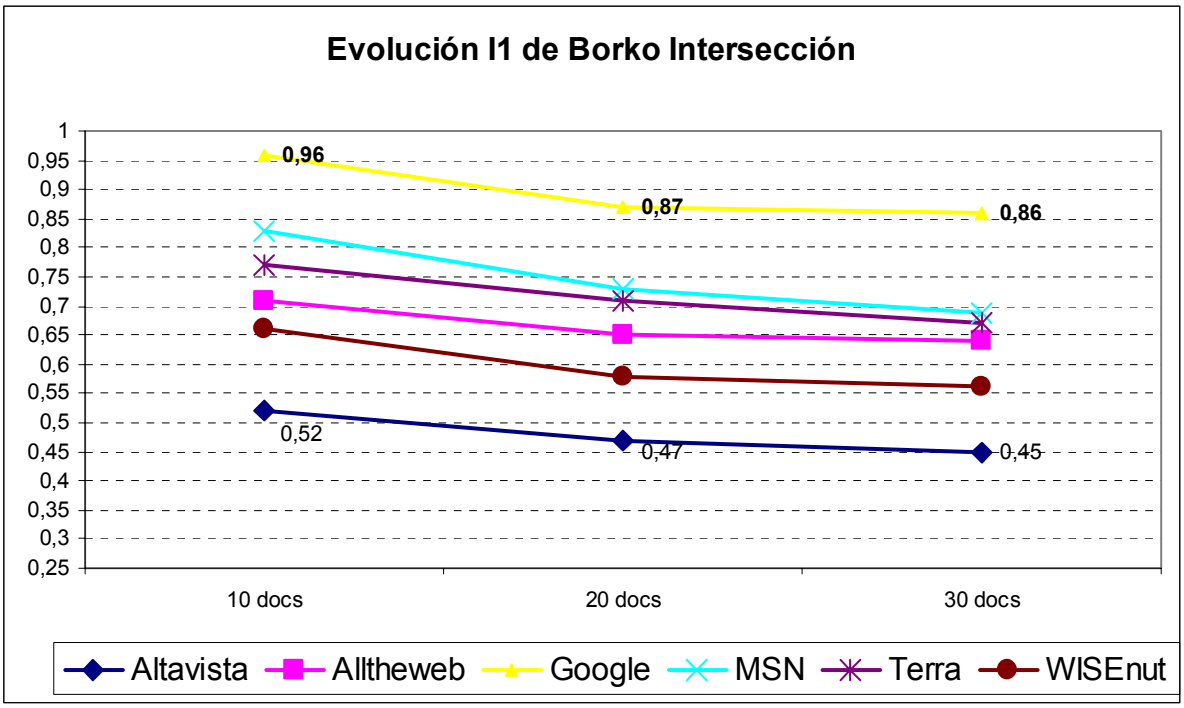


Ilustración 6.10 Evolución de la medida I1 de Borko para la operación de intersección. Fuente: elaboración propia.

Al ser los valores de la *precisión* mayores que los valores de la *exhaustividad*, es la primera de las medidas la que marca la tendencia de la evolución de la medida simple I_1 de Borko.

De este modo, la evolución se asemeja mucho a la seguida por la *precisión* (Ilustración 6.1), obteniéndose mejores valores al analizar los diez primeros documentos que cuando se analizan, sucesivamente, los veinte y los treinta primeros, aunque la diferencia entre estos dos casos es pequeña. El valor máximo que puede alcanzar la medida I_1 de Borko es *dos*, situación a la que no se acerca ningún motor por los bajos niveles de *exhaustividad*.

La Ilustración 6.11 muestra la evolución de esta medida en el caso de las operaciones combinadas de intersección y frase literal.

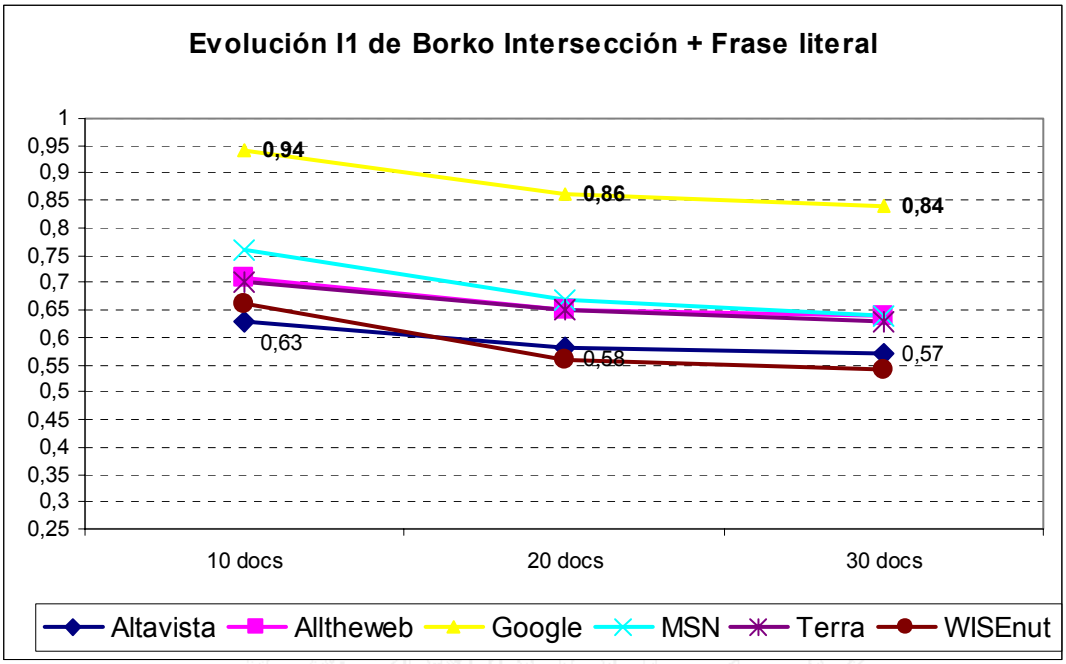


Ilustración 6.11 Evolución de la medida I1 de Borko para las operaciones combinadas de intersección y frase literal. Fuente: elaboración propia.

La *precisión* vuelve a marcar la tendencia, de similar evolución a la recogida en la Ilustración 6.3. La diferencia más significativa con respecto a las búsquedas basadas en la intersección, reside en el cierto agrupamiento de los motores, con excepción de *Google*, que obtiene valores mayores. La tabla 6.23 recoge las diferencias determinadas para la variable I_1 de Borko entre ambas modalidades de búsqueda.

	I1 de Borko		
	DIF		
	10 docs	20 docs	30 docs
Alta Vista	0,11	0,11	0,12
Alltheweb	0	0	0
Google	-0,02	-0,01	-0,02
MSN	-0,07	-0,06	-0,05
Terra	-0,07	-0,06	-0,04
WISEnut	0	-0,02	-0,02

Tabla 6.23 Diferencia de los valores de la medida I1 de Borko obtenidos en ambas modalidades de búsqueda. Fuente: elaboración propia.

Datos sobre la medida de valor simple de Borko tipificada.

Paralelamente al cálculo de la medida I_1 de Borko para cada motor en cada una de las búsquedas, se han calculado los valores de esta medida tipificados respecto al tamaño de los resultados devueltos por cada motor en cada pregunta (de distribución muy desigual). Los valores medios de esta nueva medida se recogen en la Tabla 6.24

	I1 de Borko Tipificada					
	INT			**INT + FRA**		
	10 docs	20 docs	30 docs	10 docs	20 docs	30 docs
Alta Vista	-0,98	-0,58	-0,94	-0,91	-0,14	-0,17
Alltheweb	-0,39	0,40	-0,32	-0,99	-0,50	-0,41
Google	-0,03	0,21	0,19	-0,53	0,14	0,29
MSN	0,77	0,42	0,57	-0,31	0,33	0,16
Terra	0,43	0,40	0,58	-0,42	0,27	0,21
WISEnut	0,18	-0,04	-0,07	-0,44	-0,13	0,09

Tabla 6.24 Valores medios de la medida I1 de Borko Tipificada para ambas modalidades de búsqueda. INT: búsquedas basadas en la operación de la intersección. INT + FRA: búsquedas basadas en las operaciones combinadas de intersección y frase literal. Fuente: elaboración propia.

La desigualdad del número de documentos citada, provoca que los resultados de esta medida tipificada difieran ampliamente de los obtenidos sin tipificación. La Ilustración 6.12 recoge la evolución seguida en el caso de las operaciones basadas en la intersección:

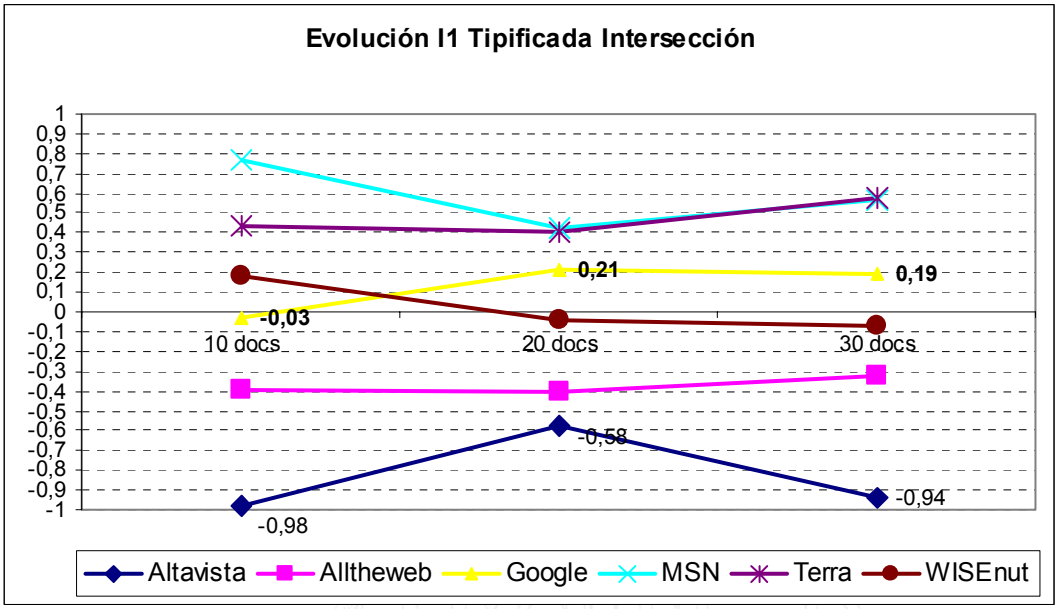


Ilustración 6.12 Evolución de la medida I1 tipificada de Borko para la operación de intersección. Fuente: elaboración propia.

Estos valores no se han sometido a contraste, ya que la medida tipificada toma como base para su cálculo valores medios previamente contrastados: exhaustividad, precisión y número de documentos devueltos, lo que hace superfluo un nuevo contraste paramétrico.

Las diferencias que presentan estos valores con respecto a los de la medida I_1 original residen básicamente en la desigual evolución de los motores. Originalmente, todos mostraban mejores resultados al analizar los diez primeros documentos y progresivamente (según se analizaban veinte o treinta documentos), descendían esos valores. En este caso, el comportamiento es muy desigual, sólo el motor *WISEnut* repite esta tendencia, mientras *MSN* muestra mejores resultados al analizar diez documentos, peores al analizar veinte documentos y mejores al analizar treinta documentos. *Terra* y *Google* (con *All the Web* en menor medida) aumentan sus valores a medida que se analizan más documentos.

Los máximos de esta función tipificada los obtiene *MSN* (muestra de diez documentos), y los motores *MSN* y *Terra* (muestras de veinte y treinta documentos). Los valores mínimos los vuelve a ofrecer *Alta Vista*.

El motor *Google* parece no explotar suficientemente el alto número de documentos como base de mejora de su pertinencia. El comportamiento de esta nueva medida tipificada, cuando se aplican las operaciones

combinadas de intersección y frase literal, es algo más homogéneo (Ilustración 6.13):

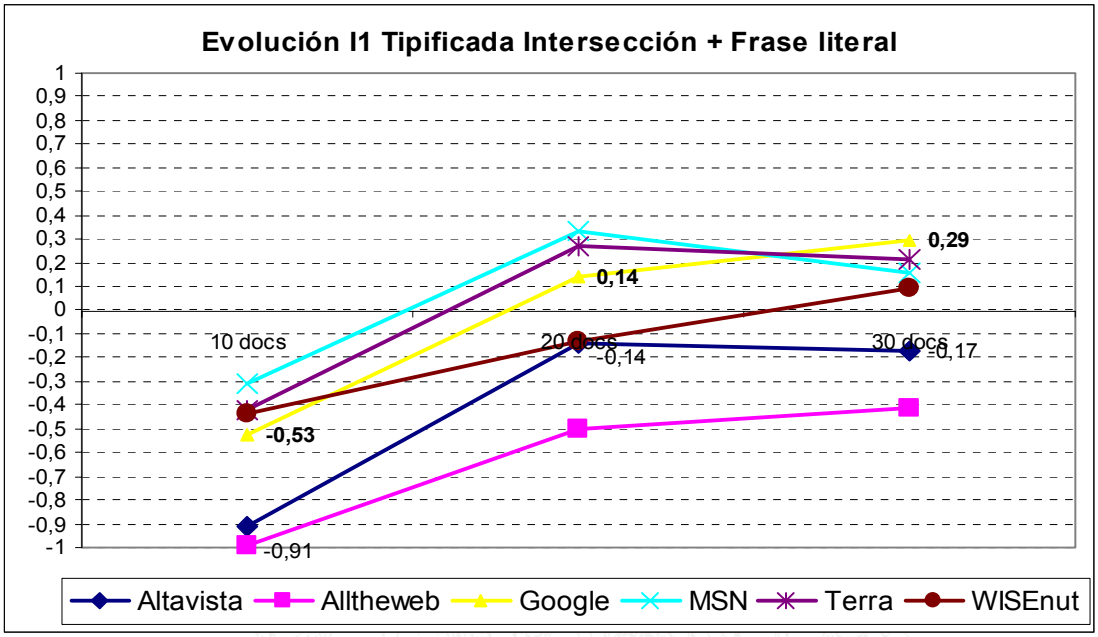


Ilustración 6.13 Evolución de la medida I1 tipificada de Borko para las operaciones combinadas de intersección y frase literal. Fuente: elaboración propia.

Con respecto a la operación de la intersección, se observa que casi todos los motores presentan una evolución positiva cuando se analizan veinte documentos frente a diez y estabilizan esa tendencia cuando se analizan treinta.

En términos de máximos valores, con esta modalidad se obtienen resultados inferiores a la anterior, aunque el motor que peores resultados presenta (*All the Web* en este caso), ofrece mejor rendimiento que *Alta Vista* en el caso anterior.

Los motores *Google* y *Terra* presentan los mejores resultados, muy similares al analizar treinta documentos y se encuentran muy próximos a los valores de *MSN* al analizar veinte y diez documentos (donde este último motor ofrece los mejores resultados). En este caso, el motor *Google* sí mantiene una cierta posición de privilegio, pero seguido muy de cerca por *MSN* y *Terra*, circunstancia que no se produce con la media I_1 original.

Comparando esta evolución con la seguida por la medida I_1 original, se advierte una tendencia totalmente contraria, la medida sin tipificar disminuye a medida que aumenta la muestra de documentos analizada, proceso radicalmente opuesto al presente.

La Tabla 6.25 recoge las diferencias de los valores de la medida I_1 tipificada en cada modalidad de búsqueda:

	I1 de Borko		
	DIF		
	10 docs	20 docs	30 docs
Alta Vista	0,07	0,44	0,77
Alltheweb	-0,6	-0,1	-0,09
Google	-0,5	-0,07	0,1
MSN	-1,08	-0,09	-0,41
Terra	-0,85	-0,13	-0,37
WISEnut	-0,62	-0,09	0,16

Tabla 6.25 Diferencia de los valores de la medida I1 de Borko tipificada obtenidos en ambas modalidades de búsqueda. Fuente: elaboración propia.

En este caso, las diferencias entre los valores de ambas modalidades de búsqueda resultan mucho más significativas que con la medida sin tipificar, siendo *Alta Vista* el que, de nuevo, mejor evoluciona, con lo cual queda claro que el volumen de documentos devueltos por cada motor sí influye en la medida de la efectividad de la recuperación de información y no siempre se aprovecha un alto número de documentos de respuesta para mejorar esos valores.

Datos sobre los agrupamientos de los motores.

A continuación se aplica el método de agrupamiento del "vecino más cercano" a las distancias existentes entre los seis motores de búsqueda, cuando se analizan diez, veinte y treinta documentos respectivamente. Para ello se han calculado las similitudes medias obtenidas en los treinta casos analizados en el experimento, tanto en la modalidad de búsquedas basadas en la intersección como cuando se combina la intersección con las búsquedas por frase literal. Se ha comprobado que estos valores medios representan verdaderamente a la distribución de los valores individuales por medio de la aplicación de la prueba de "T de Student", y posteriormente se ha intentado determinar una serie de agrupamientos conducentes a identificar grupos de motores más afines unos a otros, o bien, cuál de los motores se encuentra más alejado del resto.

Agrupamientos con muestras de diez documentos.

La Tabla 6.26 recoge las distancias medias que se han determinado para los seis motores de búsqueda con la operación de intersección y una muestra analizada de diez documentos:

	All the Web	*Google*	*MSN*	*Terra*	*WISEnut*
Alta Vista	0.87	0.90	0.95	0.95	0.88
All the Web		0.80	0.90	0.88	0.84
Google			0.85	0.83	0.85
MSN				0.70	0.87
Terra					0.85

Tabla 6.26 Distancias medias entre los motores con los resultados obtenidos por la operación de intersección y analizados diez documentos. Fuente: elaboración propia.

El algoritmo de agrupamiento aplicado sobre las distancias medias de la tabla anterior, proporciona los siguientes agrupamientos (Tabla 6.27):

Num	**Motores**	**Dist**
1	*MSN-Terra*	0.70
2	*All the Web-Google*	0.80
3	*MSN-Terra-All the Web-Google*	0.83
4	*MSN-Terra-All the Web-Google-WISEnut*	0.84
5	*MSN-Terra-All the Web-Google-WISEnut-Alta Vista*	0.85

Tabla 6.27 Agrupamientos y distancias entre los motores con los resultados obtenidos por la operación de intersección y analizados diez documentos. Num: número de agrupamiento. Dist: distancia del agrupamiento. Fuente: elaboración propia.

Los valores de distancia a los que se llevan a cabo los agrupamientos son muy altos, lo que ratifica la suposición planteada sobre la escasa similitud de los vectores de resultados. Otra hipótesis verificada es la planteada sobre la mayor distancia del motor *Alta Vista* con respecto a una hipotética respuesta ideal, ya que es el último motor que se incorpora a los agrupamientos.

En el caso de las búsquedas con operadores de intersección y frase literal, se repiten prácticamente las mismas circunstancias ya que los valores de la tabla de distancias medias obtenidas en el experimento son muy similares a los obtenidos con la operación de la intersección, tal como muestra la Tabla 6.28:

	All the Web	*Google*	*MSN*	*Terra*	*WISEnut*
Alta Vista	0.86	0.81	0.85	0.87	0.85
All the Web		0.88	0.88	0.93	0.82
Google			0.85	0.83	0.76
MSN				0.76	0.80
Terra					0.88

Tabla 6.28 Distancias medias entre los motores con los resultados obtenidos por la operación combinada de la intersección y búsqueda por frase literal una vez analizados diez documentos. Fuente: elaboración propia.

De esta tabla de distancias se derivan los siguientes agrupamientos (Tabla 6.29):

Num	**Motores**	**Dist**
1	*MSN-Terra*	0.76
2	*Google-WISEnut*	0.76
3	*MSN-Terra-Google-WISEnut*	0.80
4	*MSN-Terra-Google-WISEnut-All the Web*	0.81
5	*MSN-Terra-Google-WISEnut-All the Web-Alta Vista*	0.82

Tabla 6.29 Agrupamientos y distancias entre los motores con los resultados obtenidos por la operación de intersección y analizados diez documentos. Num: número de agrupamiento. Dist: distancia del agrupamiento. Fuente: elaboración propia.

Respecto a la intersección, desciende ligeramente la distancia del último agrupamiento y el motor *WISEnut* se agrupa con *Google* en lugar de *All the Web*, pero el comportamiento es más o menos similar.

Agrupamientos con muestras de veinte documentos.

La Tabla 6.30 recoge las distancias medias determinadas con la intersección y tras analizar muestras de un tamaño de veinte documentos:

	All the Web	*Google*	*MSN*	*Terra*	*WISEnut*
Alta Vista	0.86	0.78	0.88	0.86	0.84
All the Web		0.88	0.94	0.94	0.89
Google			0.83	0.83	0.82
MSN				0.70	0.85
Terra					0.84

Tabla 6.30 Distancias medias entre los motores con los resultados obtenidos por la operación de intersección y analizados veinte documentos. Fuente: elaboración propia.

Los valores que se han recogido en la Tabla 6.30 resultan de similares dimensiones a los obtenidos en el caso de muestras de diez documentos, por lo que no resulta de extrañar que los agrupamientos de la Tabla 6.31 sean más o menos similares:

Num	**Motores**	**Dist**
1	*MSN-Terra*	0.70
2	*All the Web-Google*	0.78
3	*All the Web-Google-WISEnut*	0.82
4	*MSN-Terra- All the Web-Google-WISEnut*	0.83
5	*MSN-Terra- All the Web-Google-WISEnut-Alta Vista*	0.83

Tabla 6.31Agrupamientos y distancias entre los motores con los resultados obtenidos por la operación de intersección y analizados veinte documentos. Num: número de agrupamiento. Dist: distancia del agrupamiento. Fuente: elaboración propia.

En este caso, aunque el tercer agrupamiento difiere del producido en la muestra de diez documentos y el último valor de distancia es algo inferior, pero el comportamiento de los motores es similar. Las distancias medias calculadas en el caso de la operación combinada de intersección y frase literal se muestran en la Tabla 6.32.

	All the Web	*Google*	*MSN*	*Terra*	*WISEnut*
Alta Vista	0.84	0.79	0.85	0.84	0.83
All the Web		0.85	0.85	0.89	0.81
Google			0.81	0.83	0.80
MSN				0.71	0.81
Terra					0.87

Tabla 6.32 Distancias medias entre los motores con los resultados obtenidos por la operación combinada de la intersección y frase literal, analizados veinte documentos. Fuente: elaboración propia.

Estas distancias vuelven a presentar dimensiones similares a los casos anteriores y los agrupamientos derivados se pueden ver en la Tabla 6.33:

Num	Motores	Dist
1	*MSN-Terra*	0.71
2	*All the Web-Google*	0.79
3	*All the Web-Google-WISEnut*	0.80
4	*MSN-Terra- All the Web-Google-WISEnut*	0.81
5	*MSN-Terra- All the Web-Google-WISEnut-Alta Vista*	0.85

Tabla 6.33 Agrupamientos y distancias entre los motores con los resultados obtenidos por la operación combinada de intersección y frase literal, analizados veinte documentos. Num: número de agrupamiento. Dist: distancia del agrupamiento. Fuente: elaboración propia.

Aquí se presentan valores similares a los calculados en el caso de la intersección y además los agrupamientos son similares.

Agrupamientos con muestras de treinta documentos.

Esta serie calculada corresponde a muestras de treinta documentos. Los valores para la operación de la intersección están en la tabla 6.34:

	All the Web	*Google*	*MSN*	*Terra*	*WISEnut*
Alta Vista	0.85	0.77	0.88	0.85	0.83
All the Web		0.87	0.93	0.93	0.87
Google			0.81	0.80	0.80
MSN				0.70	0.82
Terra					0.83

Tabla 6.34 Distancias medias entre los motores con los resultados obtenidos por la operación de la intersección, analizados treinta documentos. Fuente: elaboración propia.

No es de extrañar que los valores obtenidos sean de dimensiones similares a los calculados al analizar muestras de diez y veinte documentos y, por lo tanto, que los agrupamientos sean prácticamente los mismos y a distancias más o menos similares.

Num	Motores	Dist
1	*MSN-Terra*	0.70
2	*All the Web-Google*	0.77
3	*All the Web-Google-WISEnut*	0.80
4	*MSN-Terra- All the Web-Google-WISEnut*	0.80
5	*MSN-Terra- All the Web-Google-WISEnut-Alta Vista*	0.85

Tabla 6.35 Agrupamientos y distancias entre los motores con los resultados obtenidos por la operación combinada de intersección y frase literal, analizados treinta documentos. Num: número de agrupamiento. Dist: distancia del agrupamiento. Fuente: elaboración propia.

Por lo tanto, el comportamiento en lo relacionado con los agrupamientos de los motores de búsqueda basados en las distancias medias de sus vectores de resultado analizados es análogo, independientemente del tamaño de la muestra analizada y de la modalidad de búsqueda empleada.

Ha quedado manifiesto que los motores muestran unas grandes distancias entre ellos, que oscilan en un intervalo de 0.70 a 0.95 según los casos. Estos altos valores de distancias existentes entre los motores, corroboran el exiguo grado de coincidencia de contenidos establecido en otras partes de este estudio y el comportamiento dispar de los distintos motores a la hora de ofrecer respuestas a sus usuarios.

Llaman especialmente la atención los motores *Terra* y *MSN*, que emplean el mismo software y comparten una parte de su colección, y que a pesar de esta serie de coincidencias se encuentran ubicados a una distancia de 0.70, valor algo inferior, con respecto al resto de los motores (casi todas son superiores a 0.80), pero excesivamente alta para el nivel de parecido que, en principio, deberían poseer ambos sistemas.

Finalmente, el motor *Alta Vista* se ha ratificado como el de mayor distancia con respecto al resto de la colección en todos los casos analizados, corroborando asimismo lo reflejado en la distribución radial de las distancias medias sobre una hipotética colección ideal (Ilustración 6.9).

A esta circunstancia negativa del motor *Alta Vista*, se lunen sus pobres valores obtenidos en la medida de valor simple I_1 de Borko (Tabla 6.22) y su desigual comportamiento con la medida I_1 tipificada (Tabla 6.24).

Reflexión general sobre el comportamiento de los motores de búsqueda analizados.

Un proceso de evaluación de los resultados obtenidos por los distintos motores de búsqueda en cada uno de los aspectos estudiados en nuestro experimento, no puede finalizar sin una adecuada reflexión sobre la serie de datos que el mismo ha proporcionado, donde se pueden extractar una serie de parámetros comunes en su comportamiento.

El primer razonamiento que se desprende de este análisis está relacionado con el, relativamente escaso, número de motores de búsqueda que han satisfecho los requisitos previos para formar parte del estudio y, al mismo tiempo, las pocas coincidencias que presentan estos sistemas en su presentación y funcionamiento. Esto ha producido que la muestra objeto de estudio haya quedado restringida a un tamaño de seis motores, cuando la idea original era desarrollar un análisis de mayores dimensiones.

El segundo de los razonamientos, engendrado a partir del análisis de los datos procedentes de los juicios de *relevancia*, identifica al motor *Google* como el de mejor comportamiento en lo relacionado con la efectividad de la recuperación de la información. Al mismo tiempo, este motor ha presentado un comportamiento más estable (es decir, que ofrece una menor dispersión en sus resultados) en todas las circunstancias analizadas, tanto de modalidad de búsqueda como de tamaño de muestra analizada, lo cual constituye una garantía de rendimiento eficiente para los usuarios y justifica el enorme éxito y popularidad de este sistema. Por tanto, este motor se ha presentado como el principal candidato a ser elegido como el mejor de los sistemas de los que disponemos en la actualidad en Internet, dentro de esta dimensión de nuestra evaluación.

Vinculada a la realización de los juicios de *relevancia* se encuentra el hecho de la mayor importancia conferida a la *precisión* frente a la *exhaustividad*, situación determinada del análisis de la medida de valor simple I_1 de Borko (que ha presentado una evolución casi similar a la seguida por la *precisión*), Esta circunstancia refleja fielmente una de las tendencias más extendidas entre los usuarios de los SRI, su predilección a la *precisión* frente a la *exhaustividad*, por lo que el método seguido ha conseguido aproximarse al usuario en este aspecto.

Otra serie de considerandos importantes, procedentes del experimento, son los porcentajes de enlaces erróneos y duplicados. Aceptando la enorme dificultad existente para mantener perfectamente actualizados los índices de estas enormes bases de datos, más si cabe en un entorno afectado de tanta volatilidad, no dejan de sorprender que alguno de los sistemas

analizados ofrezca tan pobres resultados en este punto. Incluso en el mejor de los casos, el porcentaje de error no baja del tres por ciento, cifra que se nos antoja excesivamente alta especialmente cuando se han analizado muestras de treinta documentos como máximo y se supone que el motor tiende a insertar los mejores recursos entre los primeros de su respuesta. Este aspecto debería constituir uno de las cuestiones a resolver por los desarrolladores de estos sistemas, que no parecen estar dedicándoles, especialmente en algunos casos, excesiva atención.

Aunque se ha visto que el concepto de enlace duplicado es susceptible de discusión, una reflexión más o menos similar que la anterior cabría aplicar en este caso, aunque los porcentajes obtenidos han resultado de menor cuantía.

La tipificación de los resultados de *exhaustividad* y *precisión* obtenidos, con respecto al tamaño de la colección de documentos devueltos por cada motor, ha arrojado una serie de dudas, de cierta consideración sobre el rendimiento del motor *Google*, el cual ha perdido su destacada posición de privilegio al considerar esta dimensión del estudio.

Sistemas de la naturaleza de *Terra* o *MSN* que ofrecían unos resultados de nivel intermedio en el caso de las medidas originales, ofrecen un similar, e incluso mejor, rendimiento que *Google* cuando se ha considerado la medida tipificada. Este hecho no puede dejar de calificarse como sorprendente, máxime cuando se supone que el algoritmo de alineamiento de *Google* revisa los enlaces que recibe cada página para calcular su posición en la respuesta. Un mayor volumen de documentos permitiría un análisis más exhaustivo y por lo tanto, más certero y esto no parece haberse cumplido, especialmente en el caso de las operaciones basadas en la intersección. Es por ello que este resultado abre, al igual que en el caso de la frecuencia de actualización de los índices de los motores, una serie de incógnitas que posteriores estudios habrán de intentar resolver.

Otro hecho importante es que ha resultado imposible realizar un agrupamiento de alguno de los motores de búsqueda a una distancia que pudiera ser calificada como próxima, ya que el caso más favorable de distancia (el de menor, evidentemente) oscila alrededor del setenta por ciento, valor significativamente alto. Esta alta serie de distancias medias viene a evidenciar el escaso grado de solapamiento que existe entre las colecciones de los distintos sistemas de recuperación de información en la web, adelantado ya en otra serie de estudios.

Finalmente, de todo ello podemos señalar que cabría hablar de una actuación típica de los motores de búsqueda, pero que dado a que, por un

lado, el espacio en el cual se definen las vistas lógicas de los documentos y las necesidades de información es difuso y escasamente similar (se ha verificado que la colección de documentos que forman los índices de los motores no poseen una excesiva similitud, sino más bien todo lo contrario), es por lo que se detectan las diferencias que ponen en evidencia el estudio realizado.

Tablas e ilustraciones.

Tabla 6.1 Requisitos de la propuesta de evaluación satisfechos por cada motor de búsqueda. Fuente: elaboración propia. 211

Tabla 6.2.a Temas de las treinta preguntas realizadas a los motores de búsqueda seleccionados. Fuente: elaboración propia. 216

Tabla 6.2.B Expresiones formales de las treinta preguntas realizadas a los motores de búsqueda seleccionados con el operador de la intersección. Fuente: elaboración propia.....217

Tabla 6.2.c Expresiones formales de las treinta preguntas realizadas a los motores de búsqueda seleccionados con el operador de la intersección y la búsqueda por frase literal. Fuente: elaboración propia. 218

Tabla 6.3 Precisiones medias obtenidas por los motores de búsqueda con la operación de la intersección tras la realización de las 30 preguntas. Fuente: elaboración propia. 219

Recuadro 6.1 Ejemplo de Contraste de la media obtenida para *Alta Vista* con 10 documentos analizados. Fuente: elaboración propia.....221

Ilustración 6.1 Evolución de las precisiones medias de todos los motores con la operación de intersección. Fuente: elaboración propia. 222

Tabla 6.4 Coeficiente de Variación de Pearson de las precisiones medias de todos los motores con la operación de intersección.Fuente: elaboración propia. 223

Ilustración 6.2 Evolución del Coeficiente de Variación de Pearson de las distribución de la *precisión* media de todos los motores con las operaciones combinadas de intersección. Fuente: elaboración propia. 224

Tabla 6.5 Precisiones medias de todos los motores con las operaciones combinadas de intersección y búsqueda por frase literal. Fuente: elaboración propia. 225

Ilustración 6.3 Evolución de las precisiones medias de todos los motores con las operaciones combinadas de intersección y búsqueda por frase literal. Fuente: elaboración propia.....225

Tabla 6.6 Coeficiente de Variación de Pearson de las precisiones medias de todos los motores con las operaciones combinadas de intersección y búsqueda por frase literal. Fuente: elaboración propia.....226

Ilustración 6.4 Evolución del Coeficiente de Variación de Pearson de las precisiones medias de todos los motores con las operaciones combinadas de intersección y búsqueda por frase literal. Fuente: elaboración propia. 226

Tabla 6.7 Comparación de las precisiones medias de todos los motores obtenidas con ambas modalidades de búsqueda. Fuente: elaboración propia. 227

Tabla 6.8 *Exhaustividade*s medias de todos los motores con las operación de la intersección. Fuente: elaboración propia. 228

Tabla 6.9 *Exhaustividade*s medias de todos los motores con las operaciones combinadas de intersección y búsqueda por frase literal. Fuente: elaboración propia. 230

Tabla 6.10 Diferencia de las *Exhaustividade*s medias obtenidas entre los dos tipos de búsqueda. Fuente: elaboración propia. 231

Tabla 6.11 Total de documentos localizados por cada motor en las treinta preguntas realizadas con cada modalidad de búsqueda. INT: operación intersección. INT + FRA: operación combinada de intersección y frase literal. PORC: INT dividido por INT + FRA y multiplicado por cien. MED(INT): media de documentos devueltos en operación intersección. MED (INT+FRA): media de documentos devueltos en operación combinada intersección y frase literal. Fuente: elaboración propia.....232

Tabla 6.12 Total de documentos relevantes distintos localizados por cada motor en las treinta preguntas realizadas con cada modalidad de búsqueda. INT: operación intersección. INT + FRA: operación combinada de intersección y frase literal. Fuente: elaboración propia.....232

Tabla 6.13 Porcentajes de enlaces erróneos proporcionado por cada motor en el proceso de evaluación tras la realización de las treinta búsquedas. INT: resultados con operación intersección. %INT: porcentaje con operación intersección. INT + FRA: resultados con operación combinada intersección y frase literal. %INT + FRA: porcentaje con operación combinada intersección y frase literal DIF: diferencia entre el segundo y el primer porcentaje. Fuente: elaboración propia. 233

Ilustración 6.7 Evolución de los porcentajes de enlaces erróneos proporcionados por cada motor en ambas modalidades de búsqueda. Fuente: elaboración propia. 235

Tabla 6.14 Comparación porcentajes enlaces erróneos en relación al estudio de Notess. Fuente: elaboración propia, Notess, G. Search Engine Statistics: Dead reports. [En línea]. Bozeman, MT: Notes.com, 2000. <http://www.searchengineshowdown.com/stats/dead.shtml> [Consulta: 12 marzo 2002].....235

Tabla 6.15 Porcentajes de enlaces duplicados proporcionados por cada motor en el proceso de evaluación. INT: resultados con operación intersección. %INT: porcentaje con operación intersección. INT + FRA: resultados con operación combinada intersección y frase literal. %INT + FRA: porcentaje con operación combinada intersección y frase literal DIF: diferencia entre el segundo y el primer porcentaje.Fuente: elaboración propia. 236

Ilustración 6.8 Evolución de los porcentajes de enlaces duplicados proporcionados por cada motor en ambas modalidades de búsqueda. Fuente: elaboración propia. 238

Tabla 6.16 Similitudes medias obtenidas en el experimento. SIM(INT): Similitud media en las operaciones de intersección. %AUMENTO: Porcentaje de crecimiento de la similitud media. SIM(INT + FRA): Similitud media en las operaciones combinadas de intersección y frase literal. DIF: diferencia entre las similitudes medias del segundo y el primer tipo de búsqueda. Fuente: elaboración propia. 239

Tabla 6.17 Similitudes medias obtenidas en el experimento por cada par de motores, con 10, 20 y 30 documentos analizados con la operación de la intersección. AW: *All the Web*; AV: *Alta Vista*; GO: *Google*; MS: Microsoft Network; TE: *Terra*; WI: *WISEnut*. Fuente: elaboración propia.....240

Tabla 6.18 Similitud y Distancia media de cada motor con respecto al resto; operación intersección. SIM: Similitud media. DIST: distancia media (1 – SIM). Fuente: elaboración propia. 241

Tabla 6.19 Similitudes medias obtenidas en el experimento por cada par de motores, con 10, 20 y 30 documentos analizados, con la operación combinada intersección y frase literal. AW: *All the Web*; AV: *Alta Vista*; GO: *Google*; MS: Microsoft Network; TE: *Terra*; WI: *WISEnut*. Fuente: elaboración propia. 241

Tabla 6.20 Similitud y Distancia media de cada motor con respecto al resto; operación combinada de intersección y frase literal. SIM: Similitud media. DIST: distancia media (1 – SIM). Fuente: elaboración propia. 242

Ilustración 6.9 Distribución radial de la Distancia media de los motores de búsqueda con ambas posibilidades de búsqueda. Fuente: elaboración propia. 242

Tabla 6.21 Similitudes de los vectores de resultado obtenidos tema por tema en ambas modalidades de búsqueda. INT: búsquedas basadas en la operación de la intersección. INT + FRA: búsquedas basadas en las operaciones combinadas de intersección y frase literal. Σsim: suma parcial de las similitudes de cada fila Media: subtotal y total de medias de las similitudes. Fuente: elaboración propia. 243

Tabla 6.22 Valores de la medida I1 de Borko para ambas modalidades de búsqueda. INT: búsquedas basadas en la operación de la intersección. INT + FRA: búsquedas basadas en las operaciones combinadas de intersección y frase literal. Fuente: elaboración propia. 244

Ilustración 6.10 Evolución de la medida I1 de Borko para la operación de intersección. Fuente: elaboración propia. 245

Ilustración 6.11 Evolución de la medida I1 de Borko para las operaciones combinadas de intersección y frase literal. Fuente: elaboración propia.....246

Tabla 6.23 Diferencia de los valores de la medida I1 de Borko obtenidos en ambas modalidades de búsqueda. Fuente: elaboración propia. 246

Tabla 6.24 Valores medios de la medida I1 de Borko Tipificada para ambas modalidades de búsqueda. INT: búsquedas basadas en la operación de la intersección. INT + FRA: búsquedas basadas en las operaciones combinadas de intersección y frase literal. Fuente: elaboración propia.....247

Ilustración 6.12 Evolución de la medida I1 tipificada de Borko para la operación de intersección. Fuente: elaboración propia. 248

Ilustración 6.13 Evolución de la medida I1 tipificada de Borko para las operaciones combinadas de intersección y frase literal. Fuente: elaboración propia.....249

Tabla 6.25 Diferencia de los valores de la medida I1 de Borko tipificada obtenidos en ambas modalidades de búsqueda. Fuente: elaboración propia.....250

Tabla 6.26 Distancias medias entre los motores con los resultados obtenidos por la operación de intersección y analizados diez documentos. Fuente: elaboración propia. 251

Tabla 6.27 Agrupamientos y distancias entre los motores con los resultados obtenidos por la operación de intersección y analizados diez documentos. Num: número de agrupamiento. Dist: distancia del agrupamiento. Fuente: elaboración propia. 251

Tabla 6.28 Distancias medias entre los motores con los resultados obtenidos por la operación combinada de la intersección y búsqueda por frase literal una vez analizados diez documentos. Fuente: elaboración propia. 252

Tabla 6.29 Agrupamientos y distancias entre los motores con los resultados obtenidos por la operación de intersección y analizados diez documentos. Num: número de agrupamiento. Dist: distancia del agrupamiento. Fuente: elaboración propia. 252

Tabla 6.30 Distancias medias entre los motores con los resultados obtenidos por la operación de intersección y analizados veinte documentos. Fuente: elaboración propia.....252

Tabla 6.31Agrupamientos y distancias entre los motores con los resultados obtenidos por la operación de intersección y analizados veinte documentos. Num: número de agrupamiento. Dist: distancia del agrupamiento. Fuente: elaboración propia. 253

Tabla 6.32 Distancias medias entre los motores con los resultados obtenidos por la operación combinada de la intersección y frase literal, analizados veinte documentos. Fuente: elaboración propia.....253

Tabla 6.33 Agrupamientos y distancias entre los motores con los resultados obtenidos por la operación combinada de intersección y frase literal, analizados veinte documentos. Num: número de agrupamiento. Dist: distancia del agrupamiento. Fuente: elaboración propia.....254

Tabla 6.34 Distancias medias entre los motores con los resultados obtenidos por la operación de la intersección, analizados treinta documentos. Fuente: elaboración propia.....254

Tabla 6.35 Agrupamientos y distancias entre los motores con los resultados obtenidos por la operación combinada de intersección y frase literal, analizados treinta documentos. Num: número de agrupamiento. Dist: distancia del agrupamiento. Fuente: elaboración propia.....255

BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES

7 Conclusiones.

RESUMEN: Las conclusiones recapitulan los aspectos tratados a lo largo de este trabajo, y se agrupan en torno a los siguientes temas:

- La necesidad de evaluación de los SRI inherente a su propia naturaleza.
- El amplio conjunto de medidas propuestas para ponderar su efectividad y la problemática de su aplicación, tanto en los sistemas tradicionales como en el nuevo contexto de la web.
- La dispersión y disparidad de los resultados propuestos por los trabajos previos de evaluación de los SRI en la web.
- La necesidad de contar con una propuesta global de evaluación diseñada bajo una óptica multidimensional.
- La viabilidad del modelo planteado.
- La determinación de un comportamiento tipo de los motores de búsqueda a partir de los datos observado en el experimento

Conclusiones.

1. Las evaluaciones de los SRI se encuentran estrechamente vinculadas con la investigación y el desarrollo de la recuperación de información, es la propia naturaleza de estos sistemas la que propicia esta necesidad crítica de evaluación, justo como ocurre en cualquier otro campo de trabajo que aspire a ser calificado y certificado como científico.
2. Muchos trabajos se han centrado en analizar la efectividad del acceso físico a los datos contenidos en un SRI, cuando lo que verdaderamente cobra importancia hoy en día, es analizar el comportamiento del acceso lógico a los datos, esto es, el contenido informativo de la recuperación.
3. Aunque se ha desarrollado una respetable cantidad de medidas destinadas a llevar a cabo estas evaluaciones, la mayor parte de los estudios continúan empleando los juicios basados en la *relevancia* de los documentos para valorar esta efectividad. En la práctica, resulta más adecuado hablar de *pertinencia* de los documentos que de *relevancia*, en tanto que el primero de estos términos incorpora un sentido de utilidad para el usuario final. No obstante, este tipo de juicios está afectado de ciertas dosis de subjetividad a la hora de su realización.
4. Dentro de las medidas empleadas para medir la *pertinencia* de los documentos destacan la *precisión* y la *exhaustividad*. La primera de ellas no presenta problemas para su determinación, pero la *exhaustividad* es una medida de difícil cálculo. Es por ello que se han propuesto diversos métodos para el cómputo de la efectividad de los SRI basados en estos pares de valores, que permiten una aproximación, pero nunca una valoración exacta, sino aproximada o relativa.
5. Los pares de valores de *exhaustividad* y *precisión* no ofrecen una medida exacta de la efectividad de la recuperación de información por si solos, y por ello se ha desarrollado un amplio conjunto de medidas de valor simple (muchas de ellas basadas, generalmente, en estos pares de valores), para medir esa efectividad. Estas medidas de valor simple deben ser capaces de reflejar la efectividad de la recuperación de información de forma separada de otra serie de criterios, tales como el coste o el número de documentos recuperados.

6. La necesidad intrínseca de evaluación de los SRI se ha trasladado al nuevo contexto de la web, surgiendo trabajos contemporáneos al desarrollo de los primeros directorios y motores de búsqueda.
7. La especial naturaleza de la web presenta situaciones que, en el entorno de los SRI tradicionales, resultarían inverosímiles. El problema fundamental de la web es su tamaño, heterogeneidad e inconsistencia; los recursos cambian y se multiplican y la impresionante velocidad de respuesta de los motores de búsqueda no está ligada, ordinariamente, a una efectiva recuperación de información. Se manifiestan grandes divergencias en los métodos empleados para la realización de las evaluaciones de los SRI en la web y en la extensión de los mismos.
8. Una importante cantidad de trabajos se limita a analizar los aspectos explícitos (también se podrían denominar como "externos", "formales" o "testimoniales") del SRI: amigabilidad de la interface, velocidad de respuesta, formatos de presentación, documentación existente, ayuda del sistema, etc. Todos ellos presentan consideraciones notables pero no sirven para representar, en modo alguno, la efectividad de la recuperación de información.
9. La naturaleza atípica de los SRI en la web, propicia que no puedan emplearse exclusivamente las medidas basadas en la *relevancia*, sino que son necesarias medidas específicas, tal como es el caso de las ratios de enlaces fallidos, solapamiento y acierto único, que junto con la determinación de la cobertura del SRI, conforman un segundo grupo de estudios, que podemos denominar "experimentales", ya que llevan a cabo algún tipo de ensayo antes de presentar sus conclusiones.
10. En la revisión de las evaluaciones explícitas se recogen veinte trabajos que han empleado trece características diferentes en sus análisis. El número medio de características por estudio es bajo (1.9 por trabajo). Las características más utilizadas han sido: el tamaño del índice (7 veces), las posibilidades de búsqueda de cada motor (6 veces) y la interface de usuario (5 veces). Hasta siete motores aparecen citados como los mejores según el objeto del estudio. En los estudios que analizan el tamaño del índice, los resultados solían favorecer al motor *Alta Vista*, pero, desde que los estudios incluyen el motor *Google*, es este segundo motor el que ocupa una posición de privilegio. Cuando se analizan las posibilidades de búsqueda, es *Alta Vista* el que más veces es destacado y cuando se analiza la interface gráfica de usuario no existe una tendencia común en los resultados.
11. Uno de los primeros trabajos experimentales desarrollados, y el más citado posteriormente, es el elaborado por Chu y Rosenthal quienes

adaptan los criterios tradicionales empleados en los SRI clásicos al entorno de la web. Este método valora cuatro criterios: composición de los índices, capacidades de búsqueda, ejecución de la recuperación de información y el esfuerzo del usuario.

12. El número de estudios de tipo experimental es menor que el número de estudios explícitos, aunque en ellos se han utilizado más variables, circunstancias ambas absolutamente comprensibles, al ser estos estudios más complejos en su elaboración. El número medio de variables utilizadas es de 4.21 (59 variables empleadas entre 14 estudios), duplicando así la cifra obtenida en los estudios explícitos. Cinco han sido los motores citados como mejores en las diferentes evaluaciones (dos menos que en los explícitos). Las características más utilizadas han sido: la *precisión* (12 veces), el tamaño del índice (8 veces), el número de enlaces erróneos y el grado de solapamiento (5 veces). La mayor parte de las veces que se emplea la *precisión*, se muestra a *Alta Vista* como el mejor. Con el tamaño del índice no existe una tendencia clara, ya que se llegan a proponer hasta cuatro motores distintos en ocho estudios. La presencia de *Google* en los trabajos más recientes o en los actualizados de forma periódica, suele sobrellevar su designación como el de mayor tamaño de índice. En cambio, cuando no aparecía (seguramente porque el análisis era anterior a su puesta en marcha), *Alta Vista* ocupaba esa posición de privilegio.

13. El número máximo de sistemas analizados por estos trabajos no suele ser mayor de seis (circunstancia que se ha dado en 9 de 14 ocasiones). En cuanto al número de parámetros medidos en cada estudio, el de Chu y Rosenthal (curiosamente el primero) emplea hasta nueve, dedicándose la mayor parte de los restantes trabajos (un total de 7 sobre 12) a analizar cinco o seis características. Lo normal es que este grupo de evaluaciones incluya un conjunto reducido de motores, aunque no existe una tendencia clara en este punto, ya que su cobertura es muy diversa.

14. El diseño de una función de *precisión* que considera la posición de los documentos en la respuesta, propuesta por Leighton y Srivastava, y que favorece la presencia de documentos relevantes entre los primeros, aporta la novedad de integrar al cálculo de la *precisión* la estimación del funcionamiento de su algoritmo de alineamiento. Así, esta función, sin dejar de ser una medida basada en la *relevancia*, se ofrece como una medida orientada hacia el usuario.

15. Una de las últimas autoras presentadas, Ljoslund, sugiere la necesidad de hacer uso de varias escalas para ponderar el grado de *relevancia* de un documento, reconociendo así una de las situaciones más

normales que se le pueden presentar a un usuario. Aunque esta postura puede introducir algunas dosis adicionales de subjetividad en los juicios de *relevancia*.

16. Donde realmente Ljoslund resulta innovadora, es en el empleo de una función de similitud (la función de *Jaccard*) para determinar en qué porcentaje las respuestas de un motor forman un subconjunto de las respuestas de otro. Con este cálculo se establece un nuevo método de determinación del solapamiento entre los índices de los motores de búsqueda, tomando como base las respuestas que estos motores llevan a cabo a las mismas preguntas, sondeando la analogía existente entre los conjuntos de documentos gestionados por estos sistemas y el orden en el que los mismos se presentan al usuario.

17. Existe también otra serie de trabajos más centrados en la definición de un marco global para la evaluación de los motores de búsqueda o en el funcionamiento de los modelos que los sustentan, que en la comparación de la efectividad de unos motores con otros. Aunque el objeto principal de estos trabajos varía sustancialmente y resulta complicado establecer una línea común entre ellos, sí coinciden en el hecho de no ofrecer como resultado qué motor es más preciso o cuál de ellos es el de mayor crecimiento.

18. De esta serie de trabajos se desprende el convencimiento de la necesidad de aplicar un enfoque multidimensional a los procesos de análisis de la efectividad de la recuperación de información en la web, donde existe un cierto grado de relación entre las dimensiones analizadas. Con el fin de obtener esta adecuada visión, han de evaluarse conjuntamente motores de búsqueda de características similares, tanto en el tamaño de sus índices como en sus posibilidades de búsqueda. Del mismo modo, se deben tener en cuenta las restricciones geográficas que, en algunos casos presentan los motores, siendo más conveniente cotejar motores sin ningún tipo de restricción. La selección del idioma también es importante, en tanto que algunos motores prácticamente no pueden ser susceptibles de evaluación en un idioma distinto del inglés.

19. Es esencial, antes de proceder a cualquier proceso de evaluación de la *relevancia* de los documentos en un experimento, fijar debidamente toda la serie de procedimientos a seguir cuando se presenten casuísticas tales como páginas de contenido similar alojadas en distintos sitios, muchas referencias al mismo sitio web (situación muy frecuente en el caso de los portales dinámicos), redireccionamientos, etc. Otra circunstancia que debe clarificarse de antemano es la escala de determinación de la *relevancia*: binaria (relevante o no relevante) o la posibilidad de emplear criterios de *relevancia* parcial.

20. La lista de direcciones que ofrece un motor como respuesta puede entenderse, dentro de una perspectiva de espacio vectorial, como un vector de direcciones que puede compararse con otra lista (o vector) producida por otro motor en respuesta a la misma pregunta. Esta comparación puede llevarse a cabo por medio de una función de similitud para la determinación de esta semejanza que nos proporcionará información sobre la composición de los índices de los motores y sobre su alineamiento. Un alto grado de similitud equivale a que entre los índices de los motores existen bastantes documentos comunes y que sus algoritmos de alineamiento han funcionado de forma análoga.

21. Se considera imprescindible definir una función discreta de ponderación de la *relevancia*, establecida en virtud de la situación del documento dentro de la respuesta, con el fin de favorecer en el proceso de evaluación a aquellos motores que ofrecen al principio de su respuesta los documentos pertinentes. De esta manera también se incide, de forma indirecta, en el valor de la similitud de los vectores de resultados.

22. El porcentaje de enlaces erróneos se convierte en el principal indicador de la frecuencia de actualización del índice del motor de búsqueda. En nuestro experimento se ha verificado que sólo los motores *Google* y *MSN* obtienen porcentajes por debajo o próximos al cinco por ciento en ambas modalidades de búsqueda. El resto se agrupa en un intervalo comprendido entre el siete y el nueve por ciento, con excepción de *WISEnut* que se dispara a valores excesivamente altos. Este comportamiento se repite las dos modalidades de búsqueda analizadas (búsquedas basadas enla intersección y búsquedas basadas en intersección combinadas con frase literal), e induce a pensar que la frecuencia de actualización de este motor es de muy bajo nivel.

23. Al tratarse el volumen de documentos devueltos por un motor y la *precisión* (o la *exhaustividad*, en su caso) de cada respuesta, dimensiones de naturaleza distinta y por lo tanto, no susceptibles de comparación, es interesante proceder a la tipificación de los valores que ofrecen cada una de estas variables por cada motor, obteniéndose así una medida tipificada, que se denominará *precisión* normalizada (o *exhaustividad* normalizada). Con estos valores normalizados, se calculará una nueva medida de valor simple I_1 de Borko (tipificada) para verificar la incidencia del volumen de documentos devueltos por cada motor para cada par de valores E-P.

24. Los motivos para el uso de estos SRI son muy diversos, por lo que la temática de las preguntas que han de formar parte del estudio

también ha de ser amplia y variada, propiciándose así un mayor acercamiento de la evaluación al contexto de la web y proporcionando unos resultados más cercanos al usuario común de Internet que al usuario especializado. En la elaboración de estas preguntas resulta vital que las mismas posean un cierto grado de formalización, en referencia a las hipotéticas necesidades de información de un usuario particular.

25. La *precisión* calculada de los motores con la operación de intersección muestra abiertamente que todos obtienen mejores resultados cuando se analizan los diez primeros documentos recuperados, descendiendo estos valores de forma progresiva, y a un ritmo más o menos similar cuando aumenta el tamaño de la muestra analizada. Los resultados obtenidos con la modalidad de búsquedas combinadas de intersección y frase literal han resultado muy similares, aunque a excepción del motor *Google*, el resto se concentra en un intervalo próximo de valores, lo que implica un rendimiento uniforme.

26. El cálculo del Coeficiente de Variación de Pearson, para ambas modalidades de operaciones de recuperación de información, refleja que los motores de mayor valor de *precisión* obtienen menores valores de este coeficiente, lo que sugiere que los motores más precisos poseen un rendimiento más estable. Por el contrario, aquellos motores de rendimiento más discreto en su *precisión* obtienen los valores más altos en este coeficiente, es decir, poseen un comportamiento más disperso.

27. La evolución de los valores obtenidos del Coeficiente de Pearson cuando se analizan diez, veinte o treinta documentos en la modalidad de recuperación basada en la intersección, presenta un comportamiento común: existe una mayor dispersión cuando se han analizado los diez primeros documentos, atenuándose esta disgregación cuando se han analizado treinta documentos. En la otra modalidad de búsqueda, se presenta una tendencia de ligero crecimiento.

28. No se ha detectado una variación apreciable, en términos de media absoluta de todos los motores, de los niveles de *precisión* obtenidos por cada uno de ellos según las dos modalidades de búsqueda utilizadas, con excepción del motor *Alta Vista* que mejora sus resultados alrededor del veinte por ciento en la modalidad de búsquedas combinadas de intersección y frase literal.

29. En el caso de la *exhaustividad* de las operaciones basadas en la intersección se presenta nuevamente un comportamiento común de los motores, ya que todos muestran resultados similares cuando la muestra es de un tamaño de 10 y de 30 documentos, obteniéndose

valores inferiores en el caso de muestras de 20 documentos. En la otra modalidad de búsqueda no existe prácticamente ninguna diferencia entre los valores obtenidos.

30. El Coeficiente de Variación de Pearson de la distribución de las *exhaustividades* medias de cada motor, muestra que los resultados obtenidos por el motor *Google* lo convierten en el motor de menor dispersión (el de comportamiento más estable), aunque, a medida que ha ido aumentando el número de elementos de la muestra, todos los motores (excepto *Alta Vista*) se han agrupado en torno a un mismo valor, evolución que no se presentaba en el caso de la *precisión*.

31. El número de documentos recuperados por los distintos motores de búsqueda con la operación de la intersección es muy superior, al número de documentos recuperados con la segunda modalidad de búsqueda (esta modalidad es menos exhaustiva). En cambio, esta diferencia no es tan grande si únicamente se calcula el total de documentos relevantes distintos calculados por los seis motores en cada caso. En este segundo caso, las cifras se asemejan mucho más (1911 documentos frente a 1796), lo que induce a pensar que, independientemente de la modalidad de búsqueda empleada, los algoritmos de alineamiento de los motores tienden a mostrar entre los primeros treinta documentos devueltos, un conjunto de documentos relevantes de dimensiones similares.

32. El cálculo de los enlaces duplicados indica que casi todos los motores poseen porcentajes entre el uno y el tres por ciento, siendo la excepción el motor *Terra* que fija sus valores entre el cinco y el seis por ciento.

33. El cálculo de la similitud existente entre los resultados ofrecidos por los motores verifican la hipótesis planteada originalmente: se obtiene un mayor grado de coincidencia a medida que aumenta el tamaño de la muestra analizada. Este crecimiento se establece en torno al 8%. Los valores medios de similitud existentes entre los motores de búsqueda analizados alcanzan cotas inferiores al 20%, lo que explica la amplia divergencia de resultados que suelen ofrecer. La modalidad de búsqueda empleada no afecta a los valores de similitud que permanecen casi constantes.

34. Los valores medios de las similitudes obtenidas para cada par de motores en la operación de intersección apenas varían al modificar el tamaño de la muestra, manteniéndose en unas dimensiones muy constantes. Los motores más similares (por lo tanto, más próximos en un espacio vectorial), son los motores *MSN* y *Terra*, circunstancia lógica ya que comparten un porcentaje elevado de la colección y emplean el mismo motor: Inktomi.

35. Cuando se analizan las similitudes con la operación combinada de intersección y frase literal, la distribución de valores de las similitudes calculadas motor a motor resulta algo más variada (en algunos casos), con respecto a los resultados obtenidos con la operación de la intersección. Los motores más similares vuelven a ser *MSN* y *Terra*, con un porcentaje ligeramente superior.

36. Si se agruparan las distancias obtenidas con ambas modalidades de búsqueda, en un gráfico radial establecido en torno a un hipotético centro de gravedad (que podría considerarse como el núcleo de la colección de documentos de la web), los motores se encuentran más cerca de encontrar ese centro en el caso de la segunda modalidad de búsqueda, aunque las distancias obtenidas son ciertamente considerables. Con la operación de intersección, el motor *Google* sería el más cercano a la misma, seguido muy de cerca por los motores *Terra* y *MSN*. Con la segunda posibilidad de búsqueda, el motor *MSN* y el motor *Terra* (por este orden), superan a *Google*. En ambos casos, los resultados más distantes los obtiene *Alta Vista*.

37. Se ha verificado el grado de semejanza de los resultados que ofrecen los seis motores de búsqueda en cada uno de los treinta temas analizados, cotejando los resultados obtenidos en cada modalidad de búsqueda empleada, y su valor oscila alrededor del cincuenta por ciento. Como anteriormente se había calculado que las dimensiones de las colecciones de documentos pertinentes eran de tamaño similar, estos valores permiten deducir que la mitad de los documentos relevantes mostrados en su conjunto, por los seis motores analizados, suele ser similar, independientemente del tipo de búsqueda empleado.

38. El motor *Google* ha obtenido los mejores valores de la medida de valor simple I_1 de Borko, en ambas modalidades de búsqueda. Como los valores de la *precisión* son mayores que los valores de la *exhaustividad*, es la primera de las medidas la que marca la tendencia de la evolución de la medida I_1. Así, la evolución de esta medida de valor simple se asemeja mucho a la seguida por la *precisión*.

39. Cuando se ha calculado la medida I_1 tipificada con respecto al volumen de documentos devueltos como resultado (en la operación de intersección), la desigualdad de este volumen provoca que los resultados de esta nueva medida de valor simple difieran sustancialmente de los obtenidos antes de aplicar la tipificación. Las diferencias que se presentan con respecto a la medida I_1 original residen básicamente en la desigual evolución de los distintos motores. Con la medida original todos mostraban mejores resultados al analizar los diez primeros documentos y progresivamente (según se analizaban

los veinte o treinta documentos primeros), descendían sus valores. En este caso, el comportamiento es muy desigual, sólo el motor *WISEnut* repite esta tendencia, mientras *MSN* muestra mejores resultados al analizar diez documentos, desciende al analizar los veinte documentos y aumenta de nuevo su resultado a los treinta documentos. Los motores *Terra* y *Google* (con *All the Web* en menor medida) aumentan sus valores a medida que se analizan más documentos.

40. Los máximos de esta función tipificada los ha obtenido el motor *MSN* (muestra de diez documentos) y *MSN* y *Terra* (muestras de veinte y treinta documentos). Los valores mínimos los vuelve a ofrecer *Alta Vista*. Sorprende negativamente que el motor *Google* no explote suficientemente el alto número de documentos que ofrece como respuesta para aumentar sus niveles de *pertinencia*.

41. El comportamiento de esta nueva medida tipificada al aplicar las operaciones combinadas de intersección y frase literal es algo más homogéneo. Casi todos los motores presentan una evolución positiva cuando se analizan veinte documentos frente a diez y estabilizan esa tendencia cuando se analizan treinta. En términos de máximos valores, con esta modalidad se obtienen resultados inferiores a la anterior, aunque el motor que peores resultados presenta (*All the Web* en este caso), ofrece mejor rendimiento que *Alta Vista* en el caso anterior.

42. Los motores *Google* y *Terra* presentan los mejores resultados al analizar treinta documentos y se encuentran muy próximos a los valores obtenidos por *MSN* al analizar veinte y diez documentos (donde este último motor ofrece los mejores resultados). En este caso, *Google* sí mantiene una cierta posición de privilegio, pero seguido muy de cerca por *MSN* y *Terra*, circunstancia que no se produce con la media I_1 original. Comparando ambas medidas se advierte una tendencia totalmente reversa, la medida sin tipificar disminuye a medida que aumenta la muestra de documentos analizada, proceso radicalmente contrapuesto al de la medida tipificada.

43. Al realizar los agrupamientos de los motores se comprueba que los mismos se llevan a cabo a unas distancias verdaderamente altas, lo que ratifica el postulado de escasa similitud existente entre los índices de distintos motores. Otra hipótesis verificada es la planteada sobre la mayor distancia del motor *Alta Vista* con respecto a una hipotética respuesta ideal, ya que es el último motor que se incorpora a todos los agrupamientos.

44. El comportamiento de los motores, en lo relacionado con los agrupamientos basados en las distancias medias de sus vectores de resultado, es análogo, independientemente del tamaño de la muestra analizada y de la modalidad de búsqueda empleada.

Conclusión final.

Finalmente, de todas las conclusiones anteriores, podemos manifestar **que sí es posible hablar de un comportamiento típico de los motores de búsqueda**, independientemente de los mejores o peores resultados que ofrezcan, que seguramente se deben más al espacio sobre el cual se definen las vistas lógicas de los documentos y las necesidades de información, de naturaleza difusa y escasamente similar, que a las diferencias sustanciales que pueden existir entre un sistema y otro.

BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES

8 Referencias.

RESUMEN: Las referencias de esta tesis doctoral se han realizado siguiendo escrupulosamente la norma internacional ISO 690 y su versión ISO 690-2 para publicaciones electrónicas, que constituyen un importante porcentaje de las fuentes de información empleadas en este trabajo.

BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES

Referencias.

[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]

[ARE, 1997] Arents, H. *A rated selection of Internet search tools* [En línea]. Gent: University, 1997. <http://www.lib.rug.ac.be/internet/search.html> [Consulta: 11 de noviembre 2001]

[AUL, 2000] *Curso de Estadística* [En línea]. Madrid: Imagina Networks, 2000. <http://www.aulafacil.org/CursoEstadistica/Lecc-6-est.htm> [Consulta: 18 de febrero 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]

[BHA, 1998a] Bharat, K. and Broder, A. *A technique for measuring the relative size and overlap of public web search engines*. [En línea] Military Road, Lismore NSW: Southern Cross University, 1998. <http://www7.scu.edu.au/programme/fullpapers/1937/com1937.htm> [Consulta: 6 de octubre de 2001]

[BHA, 1998b] Bharat, K. and Broder, A. *Measuring the Web*[En línea] Palo Alto, CA: Digital Systems Research Center, 1998. <http://www.research.compaq.com/src/whatsnew/sem.html> [Consulta: 6 de diciembre de 2001]

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

[BOR, 2000] Borlund, P. '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". *Revista Española de Documentación Científica*, 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 Academic Publishers, 2001.

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

[CHU, 1996] Chu, H. and Rosenthal, M. *Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology* [En línea] Asis 1996 Annual Conference Proceedings, 1996. <http://www.asis.org/annual-96/ElectronicProceedings/chu.html> [Consulta: 12 de noviembre 2001]

[COL, 1990] *¿Existe el método científico?* [En línea] México: Colegio Nacional y Fondo de Cultura Económica, 1990. <http://omega.ilce.edu.mx:3000/sites/ciencia/volumen3/ciencia3/161/htm/toc.htm> [Consulta: 14 de enero de 2002]

[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]

[FER, 2001] Fernández, H. "La naturaleza de la ciencia y el método científico" [En línea]. *Psicología y Psicopedagogía*, año II, nº 5 marzo 2001. < http://www.salvador.edu.ar/ua1-9pub02-5-01.htm> [Consulta: 12 de febrero de 2002]

[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

[GON, 2000] *Resumen de Nuevas técnicas de búsqueda en Internet* [En línea] Madrid: Facultad de Informática, Universidad Complutense, 2000. <http://www.fdi.ucm.es/asignaturas/ssii_sup/SI/ResumenNuevasTecnicas_jorge.htm> [Consulta: 14 de febrero de 2002]

[GON, 2001] González, A. *Searching for Google's Successor*. [En línea] San Francisco, CA : Wired News, 1999.

<http://www.wired.com/news/technology/0,1282,45905,00.html> [Consulta: 21 de noviembre 2001]

[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.

[GWI, 1999] Gwizdka, J. and Chignell, M. 'Towards Information Retrieval Measures for Evaluation of Web Search Engines'. [En línea]. Toronto: University, Mechanical & Industrial Engineering Department, 1999. <http://www.imedia.mie.utoronto.ca/people/jacek/pubs/webIR_eval1_99.pdf> [Consulta: 12 de octubre de 2001]

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

[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 Institute, 2002. < http://inquirus.nj.nec.com/> [Comsulta: 12 de marzo de 2002]

[JOH, 2001] Johnson, F.c., Griffiths, J.R. and Hartley, R.J. *DEVISE: a framework for the evaluation of Internet search engines.* Manchester: Metropolitan University, 2001.

[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.

[LAW, 1998] Lawrence, S. and Giles, C.L. 'Searching the World Wide Web'. *Science*, vol. 280, april 1998. p.98-100.

[LEB, 1997] Lebedev, A. *Best search engines for finding scientific information in the Web*. [En línea] Moscow: State University, 1997. <http://www.chem.msu.su/eng/comparison.html> [Consulta: 10 de noviembre de 2001]

[LEI, 1995] Leighton, V. *Performance of four World Wide Web (WWW) index services: Infoseek, Lycos, WebCrawler, and WWWWorm*. [En línea]. Winona State University, 1995 <http://www.winona.msus.edu/is-f/library-f/webind.htm> [Consulta: 11 de noviembre 2001]

[LEI, 1997] Leighton, V. and Srivastava, J. *Precision among World Wide Web Search Services (Search Engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos.* [En línea] Winona State University, 1997 <http://www.winona.msus.edu/library/webind2/webind2.htm> [Consulta: 9 de diciembre de 2001]

[LEI, 1999] Leighton, V. and Srivastava, J. 'First 20 Precision among World Wide Web Search Services (Search Engines).' Journal of the American Society for Information Science 50 (July 19, 1999): 870-881. También disponible [En línea] Winona State University, 1999. <http://www.winona.msus.edu/library/webind.htm> [Consulta: 11 de noviembre 2001]

[LOJ, 1999] Lojslund, M. *Evaluation of Web search engines and the search for better ranking algorithms*. [En línea]. Trondheim: Sør-Trøndelag University, 2000. <http://www.aitel.hist.no/~mildrid/dring/paper/SIGIR.html> [Consulta: 21 de noviembre de 2001]

[LOJ, 2000] Lojslund, M. *Evaluation of twenty Web search engines on ten rare words ranking algorithms.* [En línea]. Trondheim: Sør-Trøndelag University, 2000. <http://www.aitel.hist.no/~mildrid/dring/paper/Comp20.doc>[Consulta: 21 de noviembre de 2001]

[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.

[MAL, 1998] Maldonado Martínez, A. y Fernández Sánchez, E. "Evaluación de los principales buscadores desde un punto de vista documental: recogida, análisis y recuperación de la información". *Jornadas Españolas de Documentación*, Vol. 6, 1998. p. 529-553.

[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]

[MAR, 1999] Martínez Méndez, F.J. "El salto de la gestión de información a la gestión del conocimiento". Scire, vol 5, nº 1, 1999. También accesible [En línea] <http://www.um.es/gtiweb/fjmm/elsalto.doc> Murcia: Universidad, 1999. [Consulta: 19 de octubre de 2001]

[MAR, 2000] Marcos Marín, A. "La lengua española en Internet". [En línea]. Madrid: Centro Virtual Cervantes, 2000. <http://cvc.cervantes.es/obref/anuario/anuario_00/marcos/> [Consulta: 12 de diciembre de 2001]

[MAR, 2001a] Marín Fernández, J. *Prácticas de ordenador con SPSS para windows: manual de prácticas*. Murcia: Universidad, 2001. <http://www.um.es/docencia/jmarinfe/practicas-e.pdf> [Consulta: 21 de marzo de 2002]

[MAR, 2001b] Martínez Méndez, F.J. "Aproximación general a la evaluación de la recuperación de información por medio de los motores de búsqueda en Internet". Scire, vol 6, nº 1, 2001. También accesible [En línea] <http://www.um.es/gtiweb/fjmm/ibersid2000.PDF> Murcia: Universidad, 2001. [Consulta: 19 de octubre de 2001]

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

[MIN, 2000]. Ming, H. *Comparison of Three Search Engines* [En línea]. Toronto: University, 2000.

<http://gypsy.rose.utoronto.ca/people/ming/report.html> [Consulta: 24 de noviembre de 2001]

[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]

[NET, 2001] NetRatings//Nielsen. Top 25 web properties. [En línea]. Stamford, CT: NetRatings Inc. & Nielsen Media Research, 2001. <http://pm.netratings.com/nnpm/owa/NRpublicreports.toppropertiesmonthly> [Consulta: 18 de noviembre de 2001]

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

[NOT, 1998] Notess, G. R. *Comparing Internet search engines* [En línea]. Charles Sturt University, 1999. <http://www.csu.edu.au/special/online99/proceedings99/103a.htm> [Consulta: 10 noviembre 2001]

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

[NOT, 2000b]Notess, G. R. *Search Engine Database Change Over Time*. [En línea]. Bozeman, MT: Notes.com, 2000 <http://www.searchengineshowdown.com/stats/change.shtml> [Consulta: 10 de marzo de 2002]

[NOT, 2000c] 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]

[OPP, 2000] Oppenheim, C., Morris, A. and McKnight, C. 'The evaluation of WWW search engines'. Journal of Documentation, vol 56, nº 2, March 2000. p. 190-211

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

[PET, 1997] Peterson, R.E. *Eight internet search engines compared* [En línea]. Chicago, Illinois: University, 1996-2001. <http://www.firstmonday.dk/issues/issue2_2/peterson/> [Consulta: 7 de diciembre de 2001]

[PIC, 2001] Picard, J, and Savoy, J. *Searching and Classyfing the web using hyperlinks: A logical approach*. [En línea] Neuchatel: III, Universitè, 2001. <http://www.unine.ch/info/Gi/Papers/ictai.pdf> [Consulta: 6 de diciembre de 2001]

[POR, 2000] Pors, N. O. 'Information retrieval, experimental models and statistical analysis'. *Journal of Documentation*, vol 56, nº 1 January 2000. p. 55-70

[PRE, 2000] *Programme de Recherche et d'Enseignement en Statistique Appliquée* [En línea] Bruselas: Universitè Libre, 2000. <http://www.ulb.ac.be/assoc/presta/> [Consulta: 4 de mayo de 2002]

[QUE, 1982] Quesada Paloma, V., Isidoro Martín, A. y López Martín, L.A. *Curso y ejercicios de Estadística: Aplicación a las Ciencias Biológicas, Médicas y Sociales*. Madrid: Alhambra, 1982.

[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.

[SAL, 1989] Salton, G. London: Addison Wesley, Reading, Mass, 1989.

[SAV, 2001] Savy, J. and Picard, J. 'Retrieval effectiveness on the Web'. *Information Processing and Management,* 37(4), 1991. p. 543-569.

[SCH, 2001] Schlichting, C. and Nilsen, E. *Signal Detection Analysis of WWW Search Engines*. [En línea] Seattle: Microsoft, 1997. <http://www.microsoft.com/usability/webconf/schlichting/schlichting.htm> [Consulta: 2 de diciembre de 2001]

[SEA, 2001] SearchengineWatch.com Staff. *Other Global Search Engines* [En línea]. Darien, CT: INT Media Group, 1996.

<http://www.searchenginewatch.com/links/otherglobal.html> [Consulta: 25 de marzo de 2002]

[SEA, 2002] SearchengineWatch.com Staff. *The Major Search Engines* [En línea]. Darien, CT: INT Media Group, 1996.

<http://www.searchenginewatch.com/links/major.html> [Consulta: 25 de marzo de 2002]

[SLO, 1996] Slot, M. *Web Matrix: Overview Matrix*. [En línea]. Michigan: Ambrosia Software, 1996.

<http://www.ambrosiasw.com/~fprefect/matrix/> [Consulta: 23 de noviembre de 2001]

[STO, 1996] Stobart, S. and Kerridge, S. *WWW search engine study*. [En línea]. Sunderland: University, 1996. <http://osiris.sunderland.ac.uk/sst/se/> [Consulta: 5 de noviembre de 2000]

[SUL, 2000] Sullivan , D. *StatMarket Search Engine Ratings*. [En línea]. Darien, CT: INT Media Group, 2000. <http://www.searchenginewatch.com/reports/statmarket.html/> [Consulta: 24 de noviembre 2001]

[SUL, 2001a] Sullivan , D. *Search Engine Sizes*. [En línea]. Searchenginewatch, Darien, CT: INT Media Group, 2001 <http://www.searchenginewatch.com/reports/sizes.html> [Consulta: 10 de febrero de 2002]

[SUL, 2001b] Sullivan , D. *Media Metrix Search Engine Ratings* [En línea]. Darien, CT: INT Media Group, 2000. <http://searchenginewatch.com/reports/mediametrix.html> [Consulta: 9 de noviembre 2001]

[SUL, 2001c] Sullivan , D. *Nielsen//NetRatings Search Engine Ratings* [En línea]. Darien, CT: INT Media Group, 2000. <http://searchenginewatch.com/reports/netratings.html> [Consulta: 24 de noviembre 2001]

[SUL, 2001d] Sullivan, D. *Searches Per Day*. [En línea]. Darien CT: INT Media Group, 2001 <http://www.searchenginewatch.com/reports/perday.html> [Consulta: 12 de noviembre de 2001]

[THO, 2002] Thomas, S. *5 popular search tools battle for the medal* [En línea]. San Francisco CA, CNET.com, february 2002. <http://www.cnet.com/software/0-352105-8-8741537-1.html?tag=st.sw.352105-8-8741537-9.back2352105-8-8741537-1> [Consulta: 20 de marzo de 2002]

[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]

[UDB, 1997] *Bioestadística: Métodos y Aplicaciones* [En línea]. Málaga: Unidad de Bioestadístca de la Universidad, 1997. <http://ftp.medprev.uma.es/libro/> [Consulta: 18 de enero de 2002]

[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

[VOX, 2001] *Diccionarios.com*: [En línea] *Tus diccionarios online*. Barcelona: Spes Editorial, 2001. <http://www.diccionarios.com> [Consulta: 22 de diciembre de 2001]

[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]

[WIS, 1998] Wishard, L. *Precision Among Internet Search Engines: An Earth Sciences Case Study*. [En línea]. Pennsyilvania: State University, 1998. <http://www.library.ucsb.edu/istl/98-spring/article5.html> [Consulta: 24 de noviembre de 2001]

[ZORn 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.onlineinc.com/onlinemag/MayOL/zorn5.html> [Consulta: 15 de julio de 2001]

BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES

Resumen: La recuperación de información posee una naturaleza determinista que conlleva incertidumbre al realizar una operación de búsqueda. Han proliferado cantidad de propuestas de medida de la efectividad de los sistemas de recuperación de información (SRI). La consolidación de la World Wide Web y el continuado crecimiento del número de documentos que en ella se publican, ha propiciado la creación de los SRI más avanzados y de mayor volumen hasta hora conocidos: los motores de búsqueda. Sobre estos sistemas subyacen las dudas sobre su efectividad. Su evaluación ha sido, hasta el momento, dispersa y dispar. La dispersión procede de la poca uniformidad de los criterios empleados y la disparidad surge de la aperiodicidad de los estudios y de su diferente cobertura. Nuestra propuesta de modelo de evaluación adapta medidas empleadas anteriormente - juicios de relevancia y detección de errores y/o duplicados – e implementa una función discreta de relevancia de los documentos diseñada con base en su posición dentro de la respuesta. Finalmente, analiza la similitud de los índices de los motores. Los datos obtenidos por el experimento de verificación de este modelo, permiten identificar un comportamiento tipo de los motores de búsqueda, a pesar de lo difuso y heterogéneo que resulta el espacio donde operan.

Palabras clave: Recuperación de Información / Sistemas de Recuperación de la Información / Evaluación de la efectividad de la recuperación de la información / Motores de búsqueda / World Wide Web / Internet

Abstract: Information Retrieval gets a deterministic nature that implies certain doubts about their effectiveness. Many proposals have been developed for mesauring the effectiveness of the Information Retrieval Systems (IRS). The consolidation of the World Wide Web and the growth of published documents, have caused the development of the most advanced IRS: the web search engines. Under these systems underlie the doubts about their effectivenes. The evaluation of these IRS has been, until the moment, disperse and different. The scattering is dued to the little uniformity detected into the set of criteria used, and the disparity derives from their aperiodicity and different coverage. Our proposal adjusts and combines several measures based on the relevance of the documents with the detection of dead links and duplicates references, implementing a discrete function for the measurement of the relevance, designed in basis of the documents position into the answer. The results of our experiment, allow the possibility of identifying a standard behavior of the search engines, in spite of the diffuse and heterogeneus space where they have to work.

Keywords: Information Retrieval / Information Retrieval Systems / Information Retrieval Effectivenes Evaluation / Web search engines / World Wide Web / Internet

BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES