Rendimiento de los sistemas de recuperación de información en la web:

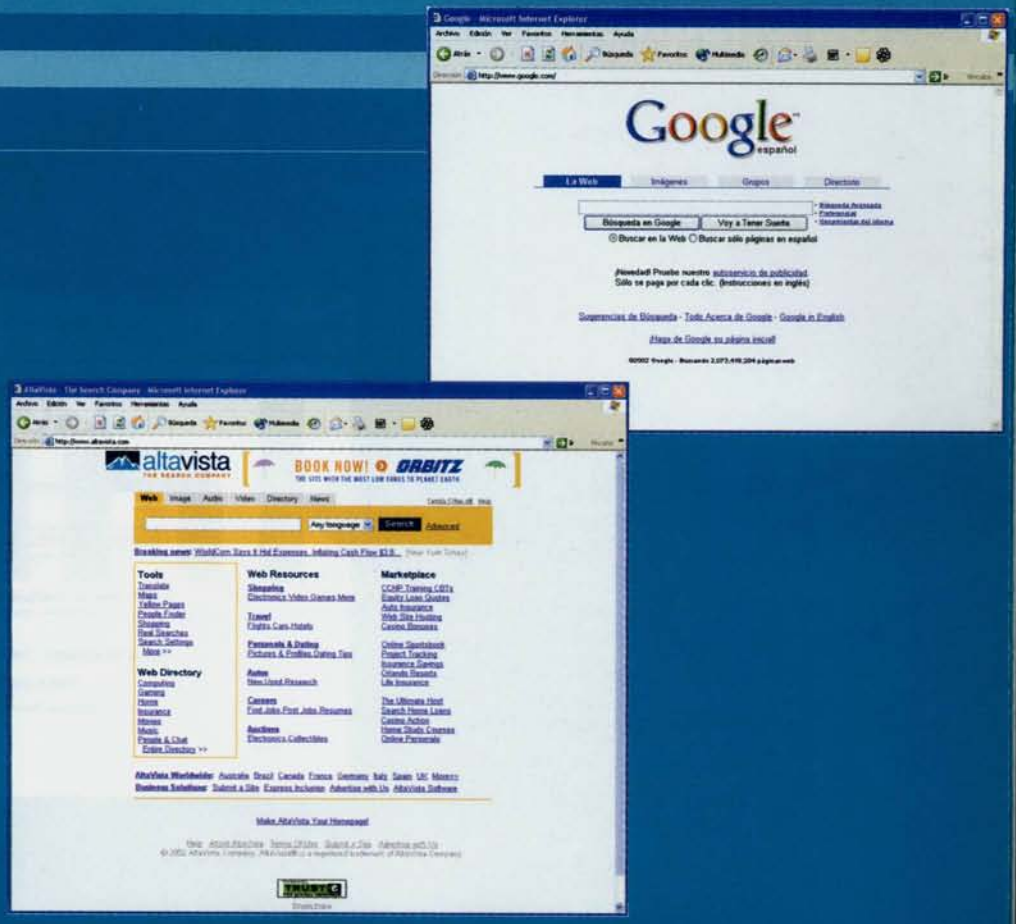
Evaluación de los servicios de búsqueda (*search engines*) Google y Altavista según consultas de los usuarios

En la década de 1990, surge la *World Wide Web* (W3), fenómeno que ha contribuido a divulgar y ampliar el uso de Internet. Ésta ha evolucionado hacia lo que podría considerarse un dinámico almacén donde albergar informaciones muy diversas en contenidos, relevancia y utilidad. Por el momento, gran parte de la responsabilidad en la búsqueda y localización de la información dispersa en la red recae en los motores de búsqueda o buscadores (Lynch, 1997).

En la web se encuentra información sobre diversos temas. Los recursos se organizan en grandes divisiones temáticas, los directorios suelen ser selectivos en la elección de los servidores que incluyen en la base de datos, siendo imposible para un servicio, abarcar toda la información disponible en la red. Por ello, un problema destacable es que reunen una proporción escasa de documentos con relación a los existentes en la W3. Los buscadores de la W3 presentan una estructura constituida por: un *robot* o *araña*, es decir, un programa que cruza la W3

moviéndose de un documento a otro, descendiendo progresivamente a través de los hiperenlaces; un *programa de indización* que indiza la información de los millones de páginas web ubicadas en servidores conectados a la red y enormes *bases de datos* a las que acceden los usuarios a través de la *interfaz* del buscador. Por tanto, los buscadores no sólo deben facilitar la localización de los recursos incluidos en sus bases de datos sino que, además, deben compilarlos.

Este mundo multimedia de gran riqueza informativa requería de nuevas herramientas de búsqueda, eficaces y sofisticadas, que permitiesen explotar todas sus posibilidades. La respuesta fueron los llamados *buscadores web o motores de búsqueda*, las "estrellas" de la localización de información en Internet.



El aumento y, en especial, la calidad de las prestaciones de los buscadores, contribuyeron desde un primer momento a hacer estas herramientas imprescindibles para los usuarios. Las continuas mejoras, hacen que hoy se hable de una nueva generación de buscadores, aunque los cambios y su evolución son constantes. Su éxito y calidad dan lugar a una gran cantidad de herramientas de consulta bien recibidas por los usuarios, como son los servicios especializados, los metabuscadores y los agentes personales de búsqueda[1].

Palabras clave: evaluación de la recuperación de información – Internet – Google – Altavista – Usuarios

Objetivos

- Evaluar dos buscadores de la World Wide Web.
- Examinar la relevancia utilizando una escala establecida.
- Analizar los datos obtenidos.
- Representar la exhaustividad (E) y la precisión (P) de los buscadores.

Materiales y métodos

A continuación se realiza un trabajo de investigación sobre la evaluación de buscadores de la World Wide Web, que constará de las siguientes etapas:

1. Seleccionar dos buscadores web.

2. Determinar las 10 preguntas a plantear.

FELQUER, Lucrecia Viviana
lfelquer@arnet.com.ar
Departamento Licenciatura en Ciencias de la Información
– Facultad de Humanidades – Universidad Nacional del Nordeste
– Avda Las Heras 727 – 3500 Resistencia – Chaco

BAZAN, Irene Olga del Valle[2]
bazanmoni@sinectis.com.ar
Departamento de Documentación - Universidad Nacional de Mar del Plata

3. Realizar las búsquedas y examinar los primeros 10 resultados recuperados.

4. Examinar la relevancia utilizando la escala establecida.

5. Proceder al análisis de datos:

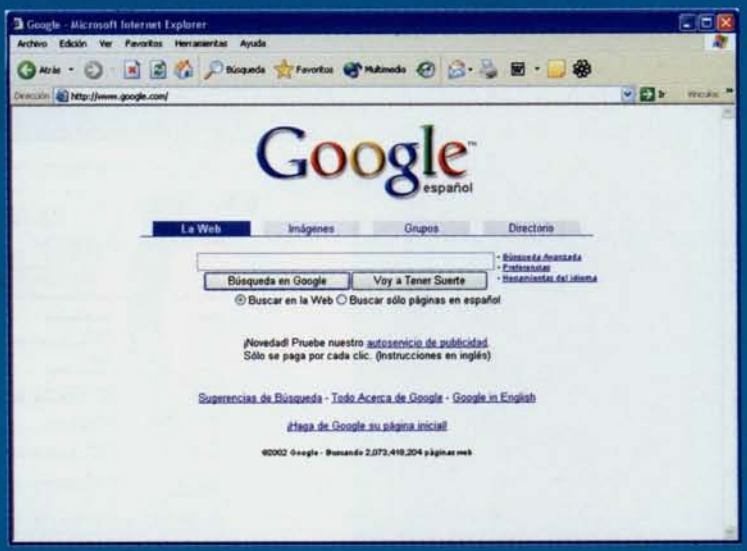
a) Ruido en la RI: n° inactivos, duplicados y relevancia 0 para todas las preguntas.

b) Promedio de resultados relevantes para las pruebas 1, 2 y 3 (de entre los diez primeros).

c) Promedio de resultados relevantes por temas de las preguntas y tipo de sintaxis (estructuradas y no estructuradas).

d) Cálculo de exhaustividad – precisión para la prueba 2.

6. Representación de exhaustividad–precisión para la prueba 2.



Los buscadores son las herramientas más importantes para localizar información en Internet. Tanto los índices temáticos como los motores de búsqueda son bases de datos de URLs.

Con el objetivo de evaluar la eficacia en la recuperación de información de los servicios de búsqueda, se analizaron dos de ellos: Altavista y Google.

Estos buscadores, fueron seleccionados al azar, aunque ambos se ajustan a los requisitos predefinidos para la realización del trabajo:

- Que fueran buscadores generales, de manera que su base de datos incluyera información sobre temas variados.
- Que fueran de carácter internacional.
- Que fueran ampliamente usados entre los usuarios y estudiosos de Internet para que la muestra fuera lo más representativa posible.

Se formularon 10 preguntas a cada uno, obteniéndose un total de 20 consultas. A continuación, se evaluó la relevancia de los 10 primeros resultados de cada consulta, dando como resultado la revisión de un total de 200 referencias.

En cada una de ellas se calculó el valor de Exhaustividad y Precisión.

El análisis efectuado demostró un mayor rendimiento en la recuperación de información por parte de Google respecto de Altavista.

Etapas de aplicación de la metodología de evaluación:

El método propuesto, consta de 5 etapas principales: 1. Planteamiento y selección de necesidades de información de los usuarios; 2. Elaboración del enunciado de búsqueda; 3. Realización de las consultas en los buscadores Altavista y Google; 4. Valoración de la relevancia de los primeros 10 resultados recuperados por cada pregunta; 5. Análisis de los resultados.

Este proceso de evaluación se inicia con la elaboración de las ecuaciones de búsqueda mediante las sintaxis correspondientes a partir de las necesidades de las información que plantean los usuarios.

1. Planteamiento y selección de necesidades de información de los usuarios

Una vez que se plantearon las demandas, se seleccionaron las preguntas que se listan más abajo, con sus respectivas sintaxis de búsqueda. La selección de las preguntas es un aspecto clave, pues de ella depende el éxito o fracaso de la recuperación. Las preguntas ofrecen el punto de partida para realizar las consultas, controlar el proceso de búsqueda, y también, para valorar los resultados que ofrece el sistema.

Las preguntas deberían presentar las siguientes características:

- Preguntas sobre las que haya información en la W3.
- Que constituyan una combinación de preguntas "fáciles" y "difíciles", con relación a la cantidad de recursos que sobre éstas pueda encontrarse.
- Que algunas sean de temas académicos y otras de ocio o comunes.
- Que se trate de preguntas heterogéneas, relacionadas con distintos temas.

En algunos casos se consideró oportuno plantear dos opciones de sintaxis de búsqueda para asegurar los resultados y conseguir de esta manera que la pregunta recupere los documentos realmente relevantes. No existe una única manera de plantear la consulta, ya que para elaborar la expresión de búsqueda, hay que decidir cuántos y cuáles términos

de la pregunta se deben incluir; además de elegir si se formula la pregunta en lenguaje natural o usando la lógica booleana. Esto da lugar a expresiones de búsqueda de diverso tipo:

- Algunas utilizan términos generales y otros más específicos.
- Algunas constan de una sola palabra y otras forman frases.
- Algunas usan la lógica booleana.
- Algunas se plantean como búsquedas de frase y otras como búsquedas en lenguaje natural.
- Algunas con nombres de persona.
- Algunas utilizan la mayúscula y el truncamiento.

En nuestro caso, en la composición de las sintaxis de búsqueda se tuvo en cuenta la naturaleza de la pregunta y las posibilidades combinatorias del álgebra booleana, del o de los términos, de la frase, etc., intentando evitar toda complejidad que pudiera entorpecer los resultados.

El lenguaje utilizado en las búsquedas fue el español, considerando el nivel de usuarios de nuestras Unidades de Información.

2. Elaboración del enunciado de búsqueda

1. Bibliotecas escolares
"bibliotecas escolares"
+ bibliotecas + escolares

2. Enseñanza de portugués
"enseñanza de portugués"
+ enseñanza + portugués

3. Colesterol

4. Juegos olímpicos Sidney 2000
"juegos olímpicos Sidney 2000"
+ juegos olímpicos + Sidney 2000

5. Reservas naturales argentinas
"reservas naturales argentinas"
+ reservas + naturales +Argentina

6. Premios Óscar 2000
"premios Óscar 2000"
+ premios + Óscar 2000

7. Premios nóbel de literatura
"premios nóbel de literatura"

8. Alimentos transgénicos
"alimentos transgénicos"

9. Historia del tenis
"historia del tenis"
+ historia + tenis

10. Cocina europea
"cocina europea"

3. Realización de las consultas en los buscadores Altavista y Google

Todas las búsquedas, se realizaron utilizando los formatos simples de recuperación que están presentes en ambos servicios. Para que el examen fuera riguroso, se formuló la misma pregunta en ambos buscadores sin que transcurriera demasiado tiempo entre el uso de los distintos motores. Otro punto importante, fue el de examinar las referencias recuperadas, accediendo al documento íntegro con la mayor rapidez posible, pues el retraso podía aumentar las probabilidades de cambio o eliminación de localización de las páginas recuperadas, y llevar a un análisis menos confiable.

4. Valoración de la relevancia de los primeros 10 resultados recuperados por cada pregunta

En este punto se determina la relevancia de cada documento recuperado. Se lo considera relevante si responde a las necesidades de información expresadas por el usuario. En general, los buscadores de la W3, ordenan los resultados en función de su relevancia respecto de la pregunta planteada. Los mejores resultados, por lo general, aparecen siempre en la parte superior de la lista de referencias.

Para evaluar la relevancia de las consultas, se accedió a cada uno de los 10 primeros resultados recuperados de un total de 200 referencias, utilizándose la siguiente escala, formada por los niveles:

Relevancia	
	0: Enlaces **duplicados, inactivos e irrelevantes** (que no satisfacen la pregunta ni recogen los términos de la ecuación de búsqueda)
	1: Enlaces técnicamente **adecuados pero no útiles** (que recogen en el HTML las diferentes partes de la pregunta pero no en el contexto adecuado o mencionan el tema en el contexto adecuado pero sólo contienen un mínimo de información relevante)
	2: Enlaces potencialmente **útiles** (que no abordan el tema en profundidad o se centran en algún aspecto específico del mismo, o páginas con al menos un enlace a otra página a la que se le asignan 3 puntos)
	3: Enlaces probablemente **más útiles** (que tratan el tema extensamente, contienen enlaces a otros documentos que tratan el tema, ofrecen una bibliografía de página web o "webibliografía")

De las 200 referencias relevantes, se obtuvieron los siguientes resultados:

Tema 1: **Bibliotecas escolares**

Buscador: **Altavista** (950 páginas)

No. de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0				1	1		1	1			4
	1						1					1
	2	1	1	1						1	1	5
	3											0

Buscador: **Google** (7.320 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0				1	1						2
	1		1						1	1	1	4
	2	1		1				1				3
	3						1					1

Tema 2: **Enseñanza de portugués**

Buscador: **Altavista** (201.860 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0		1	1		1	1					4
	1							1		1	1	3
	2	1			1							2
	3								1			1

Buscador: **Google** (941 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0	1	1		1	1	1	1	1	1		8
	1			1							1	2
	2											0
	3											0

Tema 3: **Colesterol**

Buscador: **Altavista** (7.372 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0	1		1							1	3
	1		1					1	1	1		4
	2					1	1					2
	3				1							1

Uscador: **Google** (25.000 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0				1							1
	1					1	1	1	1	1		5
	2		1	1								2
	3	1									1	2

Tema 4: **Juegos olímpicos Sidney 2000**

Buscador: **Altavista** (101.018 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0	1		1								2
	1		1			1	1		1	1	1	6
	2				1							1
	3							1				1

Buscador: **Google** (9.550 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0							1				1
	1											0
	2		1		1	1	1		1		1	6
	3	1		1						1		3

Tema 5: **Reservas naturales argentinas**

Buscador: **Altavista** (32.788 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0							1	1	1		3
	1			1							1	2
	2	1	1		1	1						4
	3						1					1

Buscador: **Google** (790 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0	1		1	1		1	1	1		1	7
	1											0
	2					1						1
	3		1							1		2

Tema 6: **Premios Óscar 2000**

Buscador: **Altavista** (14.868 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0				1	1	1	1	1	1		6
	1		1								1	2
	2	1		1								2
	3											0

Buscador: **Google** (9.540 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0			1		1						2
	1							1		1		2
	2		1		1		1		1		1	5
	3	1										1

Tema 7: **Premios nóbel de literatura**

Buscador: **Altavista** (980 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0					1	1	1		1		4
	1	1	1	1	1				1			5
	2										1	1
	3											0

Buscador: **Google** (2.520 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0				1	1				1	1	4
	1	1	1	1								3
	2						1					1
	3							1	1			2

Tema 8: **Alimentos transgénicos**

Buscador: **Altavista** (62.517 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0											0
	1				1		1			1	1	4
	2	1	1	1		1		1				5
	3								1			1

Buscador: **Google** (5.190 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0		1					1				2
	1						1		1	1	1	4
	2	1			1	1						3
	3			1								1

Tema 9: **Historia del tenis**

Buscador: **Altavista** (924.105 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0						1	1			1	3
	1	1	1	1	1	1						5
	2								1	1		2
	3											0

Buscador: **Google** (75.600 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0				1	1		1	1			4
	1						1					1
	2	1	1	1						1	1	5
	3											0

Tema 10: **Cocina europea**

Buscador: **Altavista** (294.250 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0			1	1			1		1	1	5
	1		1			1	1		1			4
	2	1										1
	3											0

Buscador: **Google** (10.400 páginas)

Nº de tema recuperado		1	2	3	4	5	6	7	8	9	10	TOTAL
Relevancia	0		1		1			1			1	4
	1	1		1			1		1			4
	2											0
	3					1				1		2

5.1 Análisis de los resultados según la relevancia de los buscadores Altavista y Google

La Figura 1 muestra para cada uno de los 10 primeros resultados obtenidos para las 10 preguntas formuladas, los inactivos, duplicados e irrelevantes, es decir lo que comúnmente se conoce como "ruido documental" en un sistema de recuperación de la información pudiendo ser considerado por los intermediarios de la información como factor negativo, no siendo así desde el punto de vista de los usuarios ya que los duplicados les muestra que la repetición de las páginas web puede demostrarles mayor relevancia. El análisis de los buscadores Altavista y Google

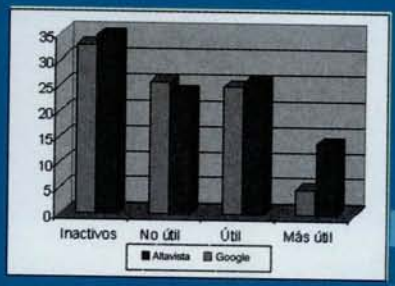


Figura 1. Resultados inactivos, duplicados e irrelevantes.

Buscadores	Relevancia			
	Inactivos (0)	No útil (1)	Útil (2)	Más útil (3)
Altavista	34	36	25	5
Google	35	25	26	14

da como resultado que ambos buscadores tienen un gran número de enlaces inactivos, duplicados e irrelevantes (34 y 35) y de enlaces técnicamente adecuados pero no útiles (36 y 25) es decir que proveen un 65% de enlaces inútiles. Al respecto, Olvera Lobos (2000), afirma que este tipo de resultados hacen pensar en la falta de actualización de las bases de datos o a un escaso índice de respuesta de algunas búsquedas, cosa que nosotros no estamos en condiciones de asumir por contar con una muestra pequeña de búsquedas. En cambio, los enlaces potencialmente útiles (25 y 26) y los más útiles (5 y 14), ocupan el 35% restante.

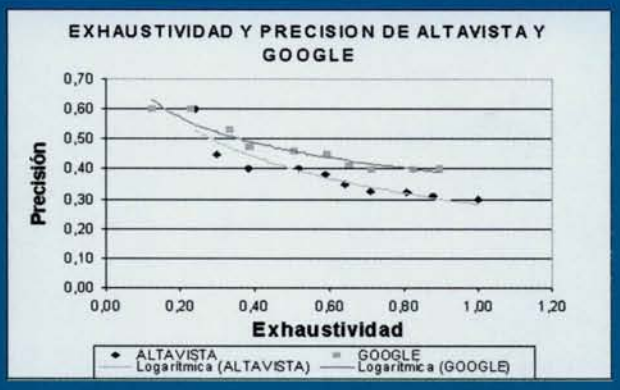


Figura 2. Representación por regresión logaritmica de la Exhaustividad y Precisión de Altavista y Google (*user-oriented recall-level average*).

	Altavista		Google	
Posición	Exhaustividad	Precisión	Exhaustividad	Precisión
1	0,24	0,60	0,12	0,60
2	0,30	0,45	0,23	0,60
3	0,38	0,40	0,33	0,53
4	0,52	0,40	0,39	0,48
5	0,59	0,38	0,51	0,46
6	0,64	0,35	0,59	0,45
7	0,71	0,33	0,65	0,41
8	0,81	0,32	0,71	0,40
9	0,88	0,31	0,83	0,40
10	1,00	0,30	0,90	0,40
	6,07	**3,84**	**5,27**	**4,73**
	Prom: 0,60	**Prom: 0,38**	**Prom. 0,52**	**Prom. 0,47**

5.2 Análisis de los resultados usando los valores medios y promedios de las medidas de exhaustividad y precisión

Para medir la exhaustividad y la precisión se tuvo en cuenta el número de temas recuperados y el número de enlaces relevantes que se recuperaron según el método de Salton y McGill para la evaluación de los SRI. Se consideraron relevantes los temas que fueron clasificados en las categorías 2 y 3 de relevancia, estando demostrado en la tabla superior.

En la Figura 2, se muestran las relaciones de exhaustividad y precisión de los buscadores Altavista y Google en donde se comprueba claramente que cuanto mayor es la exhaustividad menor es la precisión.

Conclusión

El propósito de este estudio fue el de aplicar la metodología de trabajo planteada por los autores Salton y McGill (1983), para evaluar los sistemas de Recuperación de Información en Internet. La Figura 2 refleja lo que Salton y McGill llaman el nivel medio de respuesta *(user-oriented recall-level average)* "que refleja el funcionamiento que un usuario estándar puede esperar obtener del sistema".

Es muy probable que los resultados aquí obtenidos, no reflejen exactamente la realidad actual. Lo importante es comprender que la metodología utilizada puede aplicarse de igual modo a los SRI si se presta especial atención a las características de la W3. Es por ello que hay que destacar las características esenciales de dicha metodología: 1) contar con usuarios reales que plantean preguntas reales; 2) analizar la relevancia de los 10 primeros resultados expresada en una escala de 4 grados o dígitos y 3) utilizar las medidas de exhaustividad y de precisión para evaluar las RI; y por sobretodo, podemos concluir que los resultados aquí obtenidos ponen una vez más de manifiesto que los buscadores no son sistemas inalterables de RI, es decir no son muy precisos pero sí muy exhaustivos.

Agradecimiento

A la Dra. María Dolores Olvera Lobo, docente de la carrera de Doctorado en Documentación de la Universidad de Granada de España, por la lectura crítica del trabajo y por su permanente e incondicional apoyo.

Notas:

1 Olvera Lobo, M.D. "Rendimiento de los sistemas de recuperación de información en la World Wide Web: revisión metodológica". En: *Rev. Esp. Doc. Cient.*, 23, 1, 2000.

Bibliografía

Olvera Lobo, María Dolores. "Evaluación de sistemas de recuperación de información: aproximaciones y nuevas tendencias". *El profesional de la información*. 8 (11): 4-14, nov. 1999.

——. "Métodos y técnicas para la indización y recuperación de los recursos de la World Wide Web". *Boletín de la Asociación Andaluza de Bibliotecarios*. 14 (57): 11-22, dic. 1999.

——. "Rendimiento de los sistemas de recuperación de información en la World Wide Web: revisión metodológica". *Revista Española de Documentación Científica*. 23(1): 63-77, 2000.

——. "Rendimiento de los sistemas de recuperación de información en la Web: evaluación de servicios de búsqueda (search engines)". *Revista Española de Documentación Científica*. 23(3): 303-317, 2000.

Salton, G.; Mcgill, J. *Introduction to modern information retrieval*. Nueva York: McGraw-Hill, 1983.