Análisis de la recuperación de información de motores de búsqueda y metabuscadores en la World Wide Web

Lic. Cristian Merlino Santesteban
Centro de Documentación - Facultad de Cs. Económicas y Sociales - UNMdP
Mar del Plata. Argentina
csantest@mdp.edu.ar

Palabras clave: recuperación de información, motores de búsqueda, metabuscadores, rendimiento, World Wide Web.
Keywords: information retrieval, search engines, meta-search engines, performance, World Wide Web.
Palavras-chave: recuperação da informação, motores de busca, metamotores, rendimento, World Wide Web.

OBJETIVO

Analizar la recuperación de información de motores de búsqueda y metabuscadores.

MATERIAL

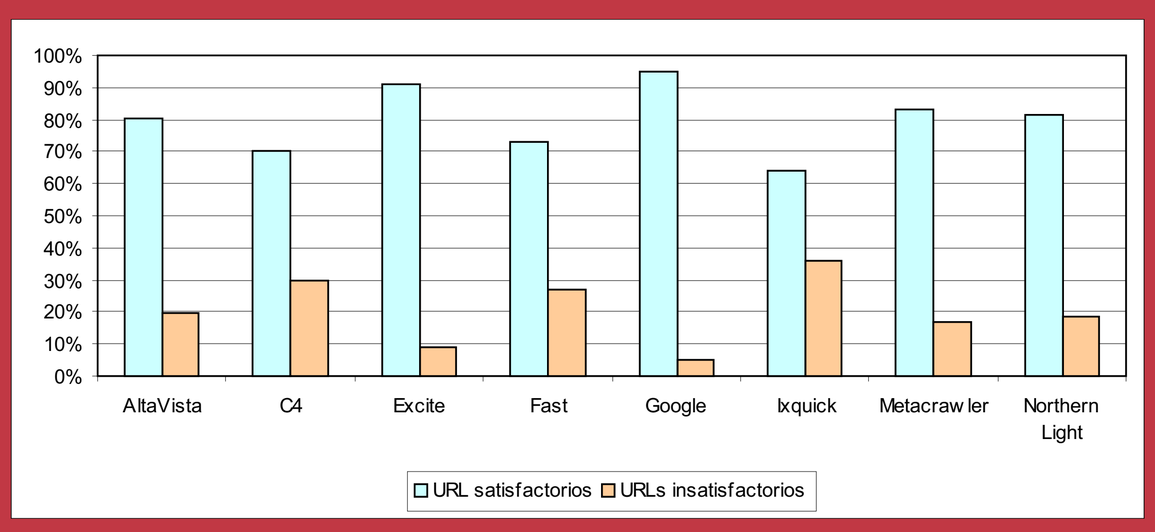
408 URLs recolectados a partir de la formulación de 10 búsquedas (*queries*) de una palabra poco frecuente, 7 en idioma español y 3 en inglés, a 8 sistemas de recuperación de información web (SRI): 5 motores de búsqueda (AltaVista, Excite, Fast, Google y Northern Light) y 3 metabuscadores (Metacrawler, C4 e Ixquick).

MÉTODO

Estudio exploratorio descriptivo. Las páginas recuperadas son categorizadas en base a una escala constituida por cinco categorías: Enlace duplicado, Enlace inactivo, Categoría cero (página irrelevante, no contiene el término), Categoría uno (página relevante potencial, contiene el término) y Categoría dos (página relevante óptima, contiene el término y su definición o explicación). Se evalua la precisión, exhaustividad, cobertura y similitud de los SRI.

RESULTADOS

La recuperación, en cantidad de aciertos, de los metabuscadores es inferior a la suma de los *ítems* recuperados por los motores que abarca y que han sido comprendidos en este estudio.
La totalidad de los servicios de búsqueda recogió más de 64% de aciertos satisfactorios (categoría uno y dos). Google (95%) y Excite (91%) obtuvieron el mayor número de registros relevantes. Mientras que Ixquick (36%) obtuvo el más alto porcentaje de URLs insatisfactorios (categoría cero, enlaces duplicados e irrelevantes) seguido por C4 (30%).



Al aplicar las categorías de relevancia en 4 pruebas, sin computar los enlaces ináctivos, se observa un gran variación en la precisión y exhaustividad de los sistemas.

Prueba 1: URLs relevantes (óptimos + potenciales), Prueba 2: URLs relevantes óptimos, Prueba 3: URLs relevantes potenciales y Prueba 4: URLs relevantes sin duplicados.

URLs / Herramienta	Relevantes potenciales		Relevantes óptimos		Relevantes (óptimos + potenciales)		Relevantes sin duplicados	
SRI	Prec.	Exh.	Prec.	Exh.	Prec.	Exh.	Prec.	Exh.
AltaVista	0,469	0,081	0,209	0,064	0,678	0,088	0,678	0,091
C4	0,416	0,135	0,122	0,180	0,538	0,135	0,466	0,128
Excite	0,170	0,022	0,180	0,028	0,350	0,027	0,350	0,029
Fast	0,619	0,206	0,204	0,125	0,823	0,185	0,762	0,198
Google	0,679	0,161	0,221	0,121	0,900	0,157	0,877	0,167
Ixquick	0,231	0,023	0,109	0,062	0,341	0,038	0,341	0,040
MetaCrawler	0,660	0,133	0,286	0,109	0,946	0,147	0,897	0,156
Northern Light	0,748	0,239	0,188	0,111	0,936	0,224	0,877	0,235

Un análisis de varianza (ANOVA) no mostró una diferencia estadísticamente significativa entre las precisiones a nivel 0,05 en la prueba 2, en tanto que en la prueba 1 (F=5,2), prueba 3 (F=3,94) y prueba 4 (F= 4,8) proporcionaron una diferencia estadísticamente significativa entre los SRI comparado con el valor crítico ($F_{(0,05;7,72)}$=2,14). El análisis de la exhaustividad de igual manera presentó una diferencia significativa en la prueba 1 (F=6,4), prueba 3 (F=7) y prueba 4 (F=7,1) . Los sistemas más precisos han sido Metacrawler (0,7), Northern Light (0,69) y Google (0,67) y los más exhaustivos Northern Light (0,19), Fast (0,17) y Google (0,15). En todas las pruebas, Excite e Ixquick, son los SRI con peor rendimiento dada a su alta cantidad de búsquedas con cero aciertos.
La tasa de solapamiento varió de 54% a 85%, cada URL fue localizado una media de 1,8 veces en los motores web y 2,7 si le adicionamos los aciertos de los metabuscadores. Los pares de buscadores presentaron coeficientes de Jaccard bastante bajos, oscilando entre 0,41 y 0,051.

CONCLUSIONES

El análisis de los URLs recolectados indica que para encontrar la mayor cantidad de registros como sea posible se debe buscar en diferentes SRI, pero el uso de metabuscadores (al menos los evaluados) no es una opción muy confiable como conjunción efectiva de varias herramientas de búsqueda.
La pertinencia de los resultados varía ampliamente de acuerdo a las diferentes categorizaciones subjetivas establecidas, y afecta en forma directa a las medidas de precisión y exhaustividad, poniendo de manifiesto el pobre rendimiento de los sistemas.
Los bajos porcentajes de solapamiento indican poca superposición entre zonas de operación de los robots en la red (reflejada en la colección de las bases de datos), y el coeficiente de Jaccard muestra el reducido grado de similitud entre los SRI, lo cual es un buen indicador para no fiarse de la respuesta de una sóla herramienta de búsqueda. La cobertura más amplia no superó el 21% de las páginas.