ESTUDIO DE USUARIOS DE DATATHÉKE. PROPUESTAS DE MEJORA UTILIZANDO EXPANSIÓN DE CONSULTAS

Zazo, Ángel F.; Berrocal, José L.; Figuerola, Carlos G.; Rodríguez, Emilio
Grupo de Recuperación Automatizada de la Información (REINA)
Facultad de Traducción y Documentación. Universidad de Salamanca
C/ Francisco Vitoria, 6-16. 37008 - Salamanca
[afzazo | berrocal | figue | aldana]@usal.es
http://reina.usal.es

RESUMEN

Uno de los aspectos más importantes a la hora de evaluar un servicio de información real es determinar la forma en que los usuarios interactúan con el sistema. El usuario plasma su necesidad informativa en un consulta y frecuentemente necesita modificarla hasta encontrar la información que considera pertinente. Uno de los mecanismos que permiten analizar el comportamiento de los usuarios es el estudio de los archivos de registro, que recogen toda su actuación. En este trabajo se presentan los resultados del análisis del archivo de registro de DATATHÉKE, uno de los servicios más conocidos del grupo de investigación REINA de la Universidad de Salamanca. A partir de dicho análisis se proponen técnicas de expansión de consultas, tanto manuales como automáticas, para mejorar la recuperación.

Palabras clave: Estudio de usuarios. Archivos de registro. Recuperación de información. Expansión de consultas.

1. INTRODUCCIÓN

Uno de los aspectos más importantes en los sistemas de recuperación de información que prestan su servicio en entornos reales es la manera en que los usuarios interactúan con el sistema. Aparte de los requerimientos del sistema para formalizar la consulta, el mayor problema consiste en determinar el conjunto de palabras o términos que expresen semánticamente esa necesidad. El problema se agrava debido al efecto de inconsistencia en la asignación subjetiva de términos a conceptos, o lo que es lo mismo, que dos personas utilizan palabras diferentes para definir los mismos conceptos [1], un

efecto muy estudiado en el campo de la indización manual [2,3]. Figuras como la sinonimia o la polisemia hacen que el mismo concepto pueda expresarse con palabras diferentes y una misma palabra pueda aparecer en documentos que tratan sobre temas distintos.

El Grupo de Investigación REINA lleva trabajando desde hace años en temas de RI, y ha puesto a disposición del público la colección documental DATATHÉKE, que puede consultarse a través de Internet utilizando el motor de recuperación Karpanta. Nuestro trabajo ha consistido en analizar los hábitos de los usuarios que utilizan nuestra servicio desde Internet. Para ello se han utilizado los archivos de registro desde noviembre de 1999 hasta la actualidad. Lamentablemente, a la hora de redactar estas líneas hemos comprobado que, debido a un fallo en uno de los programas, hemos perdido prácticamente un año de registro de la actividad sobre la colección. Ese fallo se ha ido arrastrando a lo largo de las copias de seguridad y no podemos recuperarlo. Aún así, el número de búsquedas realizadas desde noviembre de 1999 hasta septiembre de 2002, última fecha de registro, ha sido de 52.591.

Hay que señalar que solamente se registran las búsquedas e información relacionada, pero no conocemos nada de los usuarios, especialmente de su formación y de su grado de satisfacción del sistema. El estudio se ha llevado a cabo sobre la base de usuarios reales, que emplean consultas reales, con necesidades de información real, y con un motor de recuperación también real.

Este trabajo ofrece una base para la comparación de resultados en estudios similares de análisis de usuarios en OPAC y bibliotecas digitales [5,4], y, salvando las distancias, en motores de recuperación de páginas Web en Internet [7,6]. En este último caso existe una diferencia fundamental, que no es otra que la naturaleza de los usuarios que consultan el sistema. Los usuarios de Internet suponen una vasta y diferente población de personas, con muy diferentes necesidades informativas. Por el contrario, los usuarios que consultan sistemas tradicionales de recuperación de información y OPAC suelen ser usuarios con cierta formación en la búsqueda de información. Objetivos de la contribución, problemática que plantea, etc.

2. LA COLECCIÓN DATATHÉKE Y EL MOTOR KARPANTA

La colección documental DATATHÉKE contiene el vaciado de parte de los artículos publicados en más de 250 revistas y publicaciones periódicas que se

reciben en la biblioteca de la Facultad de Traducción y Documentación de la Universidad de Salamanca, especializadas en temas relacionados con la biblioteconomía, la archivística, la informática y las ciencias de la documentación. DATATHÉKE inicialmente utilizó como soporte para su difusión un BBS (*Bulletin Board System*). Aproximadamente en el año 1997 el sistema evolucionó de forma natural a prestarse en Internet. Fue entonces cuando se desarrolló la base del sistema de recuperación, al que denominamos *Karpanta* [8], para recuperar información de la colección. Hoy día es un servicio bastante conocido en entornos documentales españoles, puede consultarse en **http://milano.usal.es/dtt.htm**. Esta colección se actualiza aproximadamente tres veces al año, y en estos momentos posee unos 14.000 documentos.

Karpanta emplea el conocido modelo vectorial [10,9] utilizando lenguaje natural para la entrada de consultas. Los objetivos prioritarios eran diseñar una herramienta para la docencia y la investigación, y no un motor de búsqueda operacional, pero el resultado fue lo suficientemente robusto como para ser utilizado con éxito con pequeñas colecciones documentales. Karpanta realiza la indización y búsqueda utilizando el SGBD Microsoft Access, por su transparencia y facilidades docentes. La mayor parte de operaciones se implementan utilizando SQL bajo Visual Basic, ya que posibilita modificar de manera sencilla el cálculo de pesos o similitudes.

Volviendo a la colección DATATHÉKE, cada documento incluye el título, autor, revista, volumen, número, página, año, lengua, y también el resumen del artículo en cuestión. Hay que señalar que, salvo para el año y la lengua del artículo, el resto de información se almacena en un campo memo, esto es, la información se trata como si se dispusiera en texto libre. Este campo se procesa por el módulo de indización de Karpanta. Este módulo realiza el procesado de texto, que en nuestro sistema es bastante sencillo: en primer lugar se quitan los acentos a las vocales acentuadas y se convierten todos los caracteres a mayúsculas. En segundo lugar se eliminan las palabras vacías de contenido semántico (preposiciones, artículos, conjunciones, etc.) y aquellas que poseen una frecuencia demasiado alta en la colección documental. Sobre las palabras que quedan se aplica un proceso de lematización utilizando un S-Stemmer para unificar términos en plural. La lematización consiste en elegir convenientemente una forma para remitir a ella todas las de su misma familia

por razones de economía. Una vez obtenidos los términos índices de esta manera, se aplica el clásico mecanismo de pesado TF-IDF del modelo vectorial [11], almacenando el resultado en la base de datos.

Cuando el sistema recibe una consulta se realiza el mismo procesado léxico y se aplica el mismo mecanismo de pesado, para así disponer de representaciones homogéneas de documentos y consultas. El interfaz de consulta es sencillo e intuitivo (Figura 1). El sistema utiliza el producto escalar de los vectores que representan a documentos y consulta para obtener un ranking de documentos ordenados por similitud.

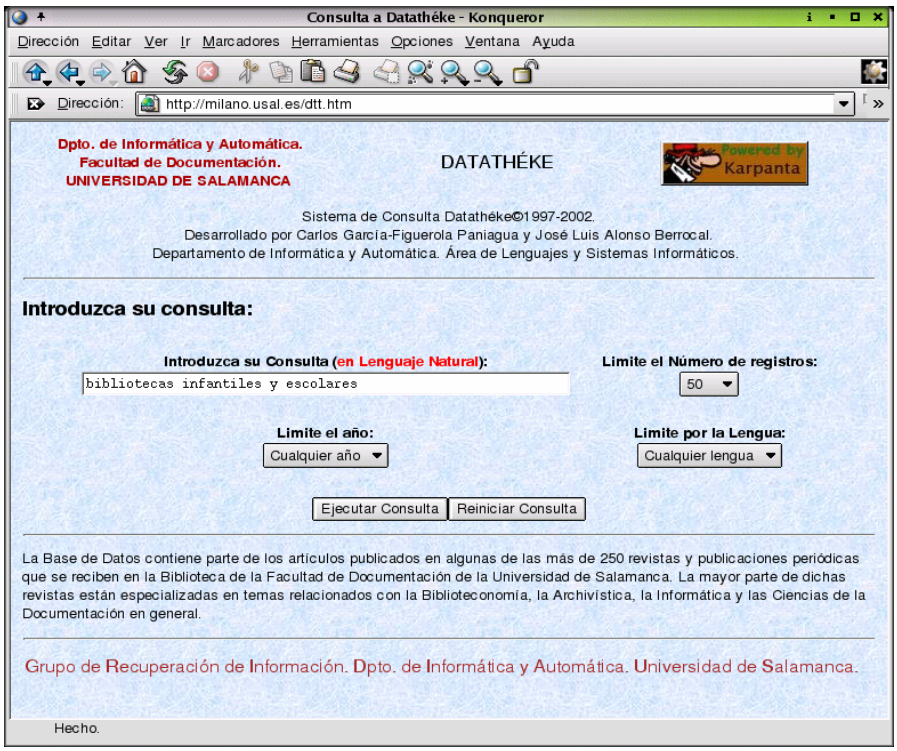

Figura 1: Interfaz de consulta para Karpanta.

3. ANÁLISIS DE DATOS

Los archivos de registro de Karpanta recogen casi toda la actividad que realiza el usuario cuando consulta el sistema. Para cada búsqueda se recoge el texto de la consulta, las opciones para año, lengua y límite de visualización de registros, y también se almacena la fecha y hora, la dirección IP y nombre por dominios del ordenador con el que el usuario se conecta. Con el actual interfaz no es posible registrar el grado de satisfacción de usuario respecto del

resultado de su búsqueda, sin embargo, podemos estimarlo si, por ejemplo, modifica su consulta en un breve intervalo de tiempo.

Dom.	País	Núm. de consultas	%	Dom.	País	Núm. de consultas	%
es	España	36.458	69,32	co	Colombia	158	0,30
IP		10.738	20,42	pe	Perú	138	0,26
net		1.786	3,40	br	Brasil	99	0,19
com		787	1,50	it	Italia	73	0,14
mx	México	551	1,05	fr	Francia	67	0,13
ar	Argentina	495	0,94	cl	Chile	63	0,12
pt	Portugal	460	0,87	uk	Reino Unido	53	0,10
ve	Venezuela	173	0,33	uy	Uruguay	48	0,09
cu	Cuba	160	0,30	otros		332	0,54

Tabla 1. Origen de las consultas.

Dominio	Organismo	Núm. de consultas	Dominio	Organismo	Núm. de consultas
usal	Univ. de Salamanca	4.350	retecal	Retecal (Operadora de CyL)	471
ttd	Telefónica Trans.Datos S.A.	3.619	uva	Univ. de Valladolid	469
ub	Univ. de Barcelona	2.384	uoc	Univ. Oberta de Catalunya	459
retevision	Retevision S.A.	2.263	ua	Univ. de Alicante	445
uv	Univ. de Valencia	1.331	uji	Univ. Jaume I	439
ucm	Univ. Complutense Madrid	1.252	udc	Univ. da Coruña	427
bne	Biblioteca Nacional	1.156	unileon	Univ. de León	395
uni2	Lince Telecomunic. S.A.	1.058	ubu	Univ. de Burgos	382
unex	Univ. de Extremadura	1.037	uclm	Univ. de Castilla-La Mancha	330
jcyl	Junta de Castilla y León	1.029	mec	Ministerio de Educacion	316
um	Univ. de Murcia	919	rediris	RedIRIS	299
us	Univ. de Sevilla	885	gva	Generalitat Valenciana	284
uc3m	Univ. Carlos III de Madrid	685	unirioja	Univ. de la Rioja	276
ugr	Univ. de Granada	542	gencat	Generalitat de Catalunya	276
usc	Univ. de Santiago	533	uco	Univ. de Cordoba	274
upv	Univ. Politecnica de Valencia	508	csic	CSIC	250

Tabla 2. Primeros subdominios para el dominio "*.es*".

3.1. ORIGEN DE LAS CONSULTAS

Podemos estimar el origen de los usuarios a partir del nombre por dominios del ordenador desde donde se lanza la búsqueda. En la Tabla 1 apreciamos que la mayor parte de conexiones proceden de España (dominio *.es*). Muchas veces no se ha podido determinar el dominio de procedencia, debido a que muchos proveedores de acceso a Internet no asignan nombre por dominios a sus usuarios, y no es posible obtener el dominio asociado. Podemos ver, por otra parte, que muchas consultas se han realizado desde Iberoamérica y Portugal.

En la Tabla 2 se muestran los subdominios dentro de España. El número más alto de búsquedas se realiza desde la propia Universidad de Salamanca; son

muchas también las búsquedas desde otras universidades. Es importante el número de consultas realizadas utilizando proveedores de acceso como Terra, Retevision, Uni2, etc. Hemos comprobado que ese número siempre ha sido alto durante todo el tiempo de recogida de datos. Esto nos indica que muchas de las consultas son realizadas por usuarios particulares desde sus hogares, pero también es cierto que en toda esta época había muchas bibliotecas y centros documentales que no contaban con una conexión permanente a la red, y muchas de esas consultas proceden de esas entidades.

3.2. USUARIOS

Es muy interesante determinar el índice de asiduidad de los usuarios, es decir, cuántas consultas se han realizado desde el mismo ordenador, y así obtener el grado de aceptación de nuestro servicio. En este sentido consideramos que el mismo usuario se conecta siempre desde la misma dirección IP. Analizando la Tabla 3 vemos que aproximadamente un tercio de los usuarios solamente se ha conectado una vez al sistema. Este dato puede ser engañoso, ya que muchos usuarios utilizan proveedores de acceso que asignan direcciones IP dinámicas. No obstante hemos preferido no eliminar las consultas procedentes de direcciones IP sin nombre o de proveedores de acceso, pues los siguientes apartados provocaría una gran distorsión en los resultados.

Núm. de consultas por usuario	Núm. de usuarios	%
1	2.960	35,47
2	1.695	20,31
3	945	11,32
4	617	7,39
5	382	4,58
6	306	3,67
7	200	2,40
8	164	1,97
9	136	1,63
10	92	1,10
+10	849	10,17

Tabla 3. Número de consultas realizadas desde la misma dirección IP.

El número total de usuarios que ha consultado el sistema, es decir, diferentes direcciones IP, ha sido de 8.346, con una media de 349,66 usuarios al mes. Un dato que indica una buena aceptación es que el 10,17% de usuarios ha realizado más de 10 consultas desde la misma localización.

3.3. OPCIONES DE BÚSQUEDA

Los usuarios que utilizan motores de búsqueda y sistemas de recuperación en Internet suelen aceptar las opciones por defecto que presenta el interfaz de consulta. En nuestro caso hemos encontrado un total de 26.001 consultas que no modifican las opciones por defecto, lo cual supone el 49,44% del total, muy por debajo de lo esperado. Ello nos indica que los usuarios que utilizan nuestro sistema tienen una buena formación en búsqueda de información. En la Tabla 4 se muestran las cifras relativas a las opciones modificadas, y su porcentaje respecto del total de consultas. Cuando se modifican las opciones por defecto, suele deberse a que el usuario desea realizar un refinamiento en su búsqueda. Volveremos sobre este aspecto en la Sección 3.7.

Opción	Núm. de consultas	%
Visualización diferente de 50 registros	9.154	19,96
Límitación por lengua	19.223	41,92
Límitación por año	11.498	25,07
Visualización diferente y limitación por lengua	5.119	11,16
Visualización diferente y limitación por año	3.188	6,95
Limitación por lengua y año	7.082	15,44
Ningún parámetro por defecto	2.105	4,59

Tabla 4. Número de consultas que no utilizan opciones por defecto.

3.4. TÉRMINOS DE LAS CONSULTAS

El número medio de términos por consulta ha sido de 2,63. Es decir, las consultas suelen ser cortas, máxime si consideramos que en el cálculo no hemos eliminado las palabras vacías. Es un valor muy parecido al de otros estudios sobre bibliotecas digitales [5] y sobre motores de búsqueda en Internet [7]. Hay que señalar que cuando las consultas son cortas la ambigüedad semántica en la que pueden incurrir aumenta notablemente respecto de consultas más largas. Efectivamente, al aumentar el número de términos en una consulta se van reduciendo las posibles ambigüedades, o lo que es lo mismo, se va especificando cada vez más la necesidad informativa del usuario.

En la Tabla 5 se muestra la distribución de consultas en función del número de términos. Podemos ver que la mayoría de consultas tienen uno o dos términos, lo cual indica que los usuarios especifican muy poco su necesidad informativa.

En tales casos hemos comprobado experimentalmente que, para la mayoría de consultas, aplicar técnicas de expansión automática de consultas mejora notablemente los resultados [12]. En la Sección 5 se realizan diferentes propuestas en este sentido.

Número de términos por consulta	Frecuencia	%
0	13	0,02
1	14.897	28,33
2	15.452	29,38
3	10.053	19,12
4	5.538	10,53
5	2.950	5,61
6	1.569	2,98
7	804	1,53
+7	1.315	2,50

Tabla 5. Número de términos por consulta.

En cuanto a los términos diferentes que se utilizan en las consultas, hay un total de 11.788 términos distintos. Para su cómputo hemos eliminado símbolos no alfanuméricos, reducido los acentos a formas no acentuadas y pasado todos los términos a minúsculas. En la Tabla 6 podemos ver los 50 términos más frecuentes. Entre los que más aparecen están las palabras vacías (de, la, en, del, el, las, a, los, sobre, etc.) que son eliminadas en el módulo de indización. En esa tabla hay que destacar otro aspecto, el de la lematización. El motor Karpanta aplica un S-Stemmer para unificar términos en singular y plural, de modo que buscar por `biblioteca' o `bibliotecas' equivale a lo mismo. Hemos observado que varios usuarios refinan sus consultas modificando términos en sentido de singular y plural, pero obtienen los mismos resultados. Más adelante analizamos las búsquedas modificadas o refinadas (Sección 3.7).

Término	Frecuencia	Término	Frecuencia
de	13.636	biblioteconomia	685
bibliotecas	4.771	recuperacion	683
la	3.447	historia	660
informacion	3.110	sistemas	648
en	2.805	bases	644
documentacion	2.189	formacion	628
biblioteca	2.011	el	612
internet	1.512	servicios	611
archivos	1.313	revistas	605
gestion	970	españa	605
universitarias	929	evaluacion	582
del	878	a	566
usuarios	872	las	554
datos	865	documental	490
referencia	822	servicio	483
bibliografia	789	bibliotecarios	480
fuentes	730	catalogacion	476
traduccion	688	los	472

Tabla 6. Términos más frecuentes en las consultas.

3.5. CONSULTAS BOOLEANAS

Ya hemos dicho que Karpanta acepta consultas en lenguaje natural; cualquier tipo de operador Booleano, de truncamiento, de posición o proximidad no es entendido por el sistema. En un porcentaje relativamente alto de consultas hemos detectado este tipo de operadores, si bien, el módulo de indización, para la mayoría de ellos, los considera palabras vacías y los elimina de la consulta. El número de consultas que utiliza alguno de estos operadores es de 4.678 (8,90% sobre el total de consultas). Es un número muy alto, esto es, son bastantes los usuarios que diseñan su consulta con unos objetivos que el sistema no puede satisfacer. En la Tabla 7 puede verse el número de consultas en las que aparecen esta clase de operadores. Para el operador IN se ha contabilizado manualmente el número de consultas, puesto que también hemos recibido consultas en inglés, en las que se utiliza esa misma palabra.

Operadores	Frecuencia	%
Y, y, AND, and, +, &	4.360	8,29
O, o, OR, or, |	265	0,50
Quot (", ')	59	0,11
NOT, not	18	0,03
Trunc. (*)	12	0,02
Selec. (IN)	6	0,01
Prox. (NEAR)	6	0,01

Tabla 7: Operadores Booleanos, de truncamiento, selección o proximidad.

3.6. SESIONES

El concepto de sesión se refiere a la serie de consultas que realiza un usuario en un intervalo corto de tiempo. Una sesión puede contener una única consulta, o varias. Hay usuarios cuyas sesiones son más largas que otras, en función del número de consultas sucesivas que lanza al sistema. La duración de la sesión la marca el usuario, pero en cualquier caso debe tener en cuenta el tiempo necesario para que el usuario pueda visualizar los registros, imprimirlos si fuera necesario, o realizar algún otro tipo de acción relacionada. Nosotros creemos que estas operaciones no deben suponer una duración más larga de 20 minutos para una consulta particular. Es decir, si un usuario realiza una consulta en un momento dado, y realiza otra consulta dentro de los 20 minutos siguientes, se trata de una consulta en la misma sesión.

Para determinar la aceptación de DATATHÉKE entre los usuarios es interesante obtener la distribución del número de consultas por sesión que se han realizado. En la Tabla 8 podemos verlo. El número total de sesiones ha sido 18.319. Cerca del 60% de las sesiones son de más de una consulta. Cuando el usuario realiza más de una consulta por sesión puede ser debido a que desea refinar más en su búsqueda, modificando los términos de la misma. Otras veces es porque desea realizar otra consulta diferente. Analizaremos en la Sección siguiente si se trata de un refinamiento de una consulta anterior. El número medio de consultas por sesión ha sido de 2,87. Curiosamente hay varias sesiones de más de 100 consultas, incluso hay una de 286 consultas.

Número de consultas por sesión	Frecuencia	%
1	7.651	41,77
2	4.219	23,03
3	2.270	12,39
4	1.348	7,36
5	842	4,60
6	568	3,10
7	373	2,04
8	258	1,41
9	195	1,06
10	120	0,66
+10	475	2,59

Tabla 8: Distribución del número de consultas por sesión.

3.7 REFINAMIENTO DE CONSULTAS

Es importante analizar el comportamiento del usuario cuando realiza más de una consulta por sesión, por si se trata de una reformulación de la consulta original para obtener mejores resultados. Es una situación bastante habitual. El usuario plantea un primera búsqueda, y a la vista de los resultados obtenidos, la va refinando cada vez más para encontrar los documentos relevantes a su necesidad informativa. Para no enmascarar resultados, hemos separado expresamente aquellas consultas iguales a la consulta anterior, ya que hemos detectado que existe un porcentaje alto de usuarios que repiten exactamente la misma consulta anterior, aunque eso sí, modificando la lengua, el año o el límite de registros a visualizar. Se trata de 11.902 consultas, es decir, el 22,63% sobre el total de consultas. Esta situación refleja un refinamiento en la búsqueda, ya que el usuario lanza la misma consulta pero con diferentes condiciones de año, lengua y/o número de registros a visualizar.

Para el resto de consultas hay que determinar el número de términos comunes en consultas consecutivas de la misma sesión, pues ello nos da pistas de si existe o no refinamiento de la búsqueda. En la Tabla 9 podemos ver que hay 13.121 consultas diferentes a la anterior. El resto, 9.250 consultas (el 17,59% sobre el total de búsquedas), tiene al menos un término común con la anterior, es decir, se trata de una consulta refinada. Si consideramos el dato anterior (11.902 búsquedas que son el refinamiento en año, lengua y/o visualización de registros de una consulta anterior), podemos concluir que el 40,22% de todas las consultas que ha recibido DATATHÉKE son reformulaciones de una consulta anterior. Es una cifra muy alta, típica de sistemas gratuitos de recuperación de información en línea.

Número de términos comunes	Frecuencia	%
0	13.121	58,65
1	3.833	17,13
2	2.386	10,67
3	1.194	5,34
4	632	2,83
5	285	1,27
6	177	0,79
7	196	0,88
8	154	0,69
9	102	0,46
10	61	0,27
+10	230	1,03

Tabla 9: Distribución del número de términos comunes en consultas sucesivas.

Solamente nos queda por analizar cómo refinan los usuarios sus búsquedas. Para el primer tipo de refinamiento, el que mantiene el texto de las consultas, hemos comprobado que principalmente se produce en los dos primeros aspectos, lengua y año, y en menor media en el límite de visualización.

Para el segundo tipo, el que modifica la consulta manteniendo términos comunes, hemos comprobado que los usuarios refinan fundamentalmente sus necesidades informativas con cambios muy pequeños: añadiendo un término, modificando otro, etc. En la Tabla 10 podemos ver el número de términos que se modifican en consultas sucesivas. Se han tomado exclusivamente las consultas que tienen algún término común, de modo que el valor de cero indica que se ha sustituido un término por otro, un valor de 1 indica que se ha añadido un nuevo término a la consulta, un valor de -1 indica que se ha eliminado un término de la consulta. Podemos observar que en la reformulación de consultas se suele dar la sustitución de un término por otro, y también la ampliación de uno o dos términos.

Número de términos añadidos	Frecuencia	%
+5	142	1,54
5	111	1,20
4	230	2,49
3	513	5,55
2	1.170	12,65
1	2.750	29,73
0	2.951	31,90
-1	704	7,61
-2	317	3,43
-3	157	1,70
-4	88	0,95
-5	57	0,62
+(-5)	60	0,65

Tabla 10. Número de términos añadidos en consultas refinadas.

4. RESUMEN Y COMPARACIÓN CON OTROS ESTUDIOS

Destacamos los siguientes resultados:

- Se han analizado 52.591 consultas realizadas desde 8.346 direcciones IP diferentes, durante el intervalo de tiempo que va desde noviembre de 1999 hasta septiembre de 2003.
- La mayoría de búsquedas se ha realizado desde el dominio "*.es*". No se ha podido determinar el dominio del 20,42% de direcciones IP, debido

fundamentalmente a que son direcciones IP dinámicas de proveedores de acceso a Internet. Esto nos indica un elevado número de conexiones desde el hogar, pero también, que muchos centros de documentación siguen teniendo conexiones no permanentes a la red.

- Un tercio de los usuarios sólo se ha conectado una vez a DATATHÉKE. Pero hay un 10,17% de usuarios que ha realizado más de 10 búsquedas. Esto significa una aceptación muy elevada de nuestro sistema.
- La mitad de los usuarios ha utilizado las opciones por defecto. Se trata de un porcentaje pequeño que indica una buena formación de los usuarios.
- Las consultas son muy cortas, con una media de 2,63 términos. Es una media que se acerca bastante a los motores de búsqueda en Internet, con una media 2,2 de términos [13], y a otras bibliotecas digitales [5], con una media de 2,43 , pero está por debajo de los sistemas de recuperación clásica, con bastante más de tres términos por consulta.
- Los usuarios han utilizado 11.788 términos diferentes. Suelen buscar sobre sistemas de información de bibliotecas y centros de documentación.
- Hay un 8,90% de consultas que utilizan algún tipo de operador Booleano, de proximidad, posición, etc. Es decir, casi una de cada diez consultas se diseña con un objetivo que el sistema no puede satisfacer.
- El número de consultas por sesión es pequeño. La media ha sido de 2,87 consultas. Casi el 60% de los usuarios ha realizado dos o más consultas en la misma sesión.
- Existe un elevado número de consultas que se reformulan para obtener mejores resultados. Un 40% de todas las consultas que ha recibido DATATHÉKE son un refinamiento de una consulta anterior. Aproximadamente la mitad se corresponden con un refinamiento por lengua y año, principalmente, y en mejor medida modificando el límite de visualización de registros. Para la otra mitad hemos comprobado que la reformulación de consultas predomina la sustitución de un término por otro, y después la ampliación de uno o más términos a la consulta original.

Para comparar nuestros resultados con otros trabajos de investigación hemos considerado el trabajo de Jones et al. [5] y de Spink et al. [13]. El primero de ellos investiga los archivos de registro para la colección Computer Science Technical Reports (CSTR) de la New Zeland Digital Library (NZDL), una

biblioteca digital accesible a través de Internet. En ese trabajo se analizan 32.000 consultas entre abril de 1996 y julio de 1997. El segundo estudio analiza el registro del buscador Excite para el 16 de septiembre de 1997, con un millón de consultas. A pesar de que parten de datos bastante antiguos, estos artículos son buenos referentes en este tipo de investigaciones.

Existen más puntos en común entre nuestro trabajo y el primero de los artículos, como era de esperar, al tratarse de dos bibliotecas digitales, con colecciones específicas de documentos, mientras que Excite indexa documentos web de muy diversa índole. Sin embargo, en contraste con nuestro estudio, esos dos artículos parten de sistemas de recuperación que admiten tanto consultas en lenguaje natural como búsqueda Booleana. Para ambos, el método de búsqueda por defecto es en lenguaje natural, en el que los resultados se ordenan de acuerdo a algún criterio interno del sistema, igual que DATATHÉKE. Pero además en los mismos también pueden realizarse búsquedas con operadores Booleanos. En este sentido, el trabajo de Jones et al. [5] muestra una utilización razonablemente elevada de operadores booleanos en torno al 20%, pero solamente el 8% para el trabajo sobre Excite. Esto nos indica que existe cierto grado de formación de los usuarios que consultan una biblioteca digital, en contraste con los usuarios que utilizan los motores de búsqueda en Internet.

En cuanto al refinamiento de consultas, hemos obtenido unos valores similares a los obtenidos en el estudio que hacen Jansen et. al. [6] al analizar aproximadamente 51.000 consultas también de Excite.

5. EXPANSIÓN AUTOMÁTICA DE CONSULTAS

En un trabajo reciente de uno de los integrantes de nuestro grupo de investigación se han analizado las técnicas de expansión en recuperación de información [12], de manera que podrían aplicarse de forma automática aquellas que mejorasen notablemente la recuperación sin necesitar de un coste computacional elevado. Son técnicas totalmente automáticas, en contraposición a otras que requieren de la presencia del usuario, como las típicas de motores de búsqueda en Internet (“páginas similares”, “*more like this*”, etc.), o la denominada realimentación de consultas con criterios de relevancia del usuario [14].

La idea es la siguiente: ya que las consultas que realizan los usuarios suelen ser muy cortas, con lo cual el nivel de ambigüedad es elevado, y por tanto, los resultados no son siempre los esperados, intentemos mejorar la consulta construyendo una nueva consulta. Se han planteado diversos mecanismos para hacerlo, pero todos ellos realizan una ampliación de nuevos términos a la consulta inicial y un recálculo de la importancia de cada término en la nueva consulta. Con esta expansión se mejora el resultado de la búsqueda, pero el coste computacional y el tiempo de respuesta aumentan.

En [12] se utiliza una colección de pruebas obtenida de DATATHÉKE, cuyas consultas poseen características similares a las que los usuarios lanzan al sistema real. Así pues, la aplicación de las técnicas de expansión desarrolladas para esa colección deberían proporcionar unos resultados similares si se aplican a la colección completa. A continuación se describen brevemente las técnicas de expansión que podrían utilizarse, así como la mejoría en los resultados que se ha alcanzado utilizando la colección de pruebas con cada uno de ellas.

5.1. PSEUDO REALIMENTACIÓN DE CONSULTAS

En la realimentación de consultas con criterios de relevancia del usuario, éste revisa los documentos que le muestra el sistema (documentos recuperados) e interviene marcando como positivos aquellos que considera relevante a su necesidad informativa, y como negativos los que considera no relevantes. El sistema entonces construye una nueva consulta teniendo en cuenta la consulta original y los documentos marcados como relevantes y no relevantes. La forma habitual de construir la nueva consulta utiliza el algoritmo de Rocchio [15]:

$$\vec{q'} = \alpha\vec{q} + \frac{\beta}{n_{rel}} \sum_{d_j \in rel} \vec{d_j} - \frac{\gamma}{n_{norel}} \sum_{d_j \in norel} \vec{d_j}$$

donde $\boldsymbol{q}$ es el vector de la consulta original, $\boldsymbol{q'}$ el vector de la consulta realimentada, $\boldsymbol{d_j}$ es el vector que representa al j-ésimo documento, $\boldsymbol{n_{rel}}$ es el número de documentos considerados relevantes por el usuario, $\boldsymbol{n_{norel}}$ es el número de documentos considerados no relevantes, y α, β y γ son parámetros para ajustar el impacto de los documentos relevantes y los no relevantes.

Pues bien, la pseudo realimentación de consultas utiliza el mismo mecanismo, pero supone que los 5, 10 ó 20 primeros documentos recuperados para una

consulta dada son relevantes, utilizando entonces los términos que aparecen en los mismos para expandir la consulta original. En general se utiliza la fórmula de Rocchio con el coeficiente $\gamma = 0$. La mayor ventaja de este mecanismo es que no necesita de la intervención del usuario, de manera que puede realizarse automáticamente en un proceso que es conceptualmente simple y de sencilla aplicación.

Utilizando la colección de pruebas hemos obtenido que los mejores resultados se producen cuando se pone α a 1,0 y β está entre 0,2 y 0,4, tomando como relevantes los 5 primeros documentos recuperados, y utilizando una consulta realimentada con entre 10 y 30 términos. En esta situación se consiguen porcentajes de mejoría en torno al 10% para la precisión media, y al 15% para la precisión a 10 documentos vistos (generalmente se denota como P@10).

5.2. TESAUROS

En Recuperación de Información un tesauro es una matriz que mide relaciones entre términos. La matriz de relación puede interpretarse como una descripción semántica de cada término, y por tanto, utilizarse en el proceso de expansión de consultas. Existen varios mecanismos para construir automáticamente un tesauro de términos, aunque la forma más sencilla es utilizando las siguientes técnicas, que más adelante desarrollaremos con más detalle:

- Medida de valores de coocurrencia: tesauros de asociación.
- Transposición de la matriz documentos-términos: tesauros de similitud.

Hay que decir que los tesauros se pueden obtener global o localmente, dependiendo de si para su construcción se utilizan todos los documentos de la colección, o solamente los primeros documentos recuperados. Nosotros hemos obtenido que los resultados son mejores con tesauros locales, aunque su construcción requiera que se realicen en tiempo real, por eso se busca que el número de documentos utilizados sea pequeño (en torno a 10 documentos).

Por otro lado, para aplicar la expansión de consultas hay que selección los mejores términos que se añadirán a la consulta original. Para ello la mejor solución es obtener un valor de relación con toda la consulta original, y no con cada término por separado (es lo que Qiu y Frei [16] denominan "*query*

concept"'). Lo habitual es, dado un término candidato, obtener su valor de relación con toda la consulta como la suma de los valores de relación con cada uno de los términos por separado.

5.2.1. Tesauros de asociación

El primer tipo consiste en medir relaciones de coocurrencia. Si dos términos aparecen frecuentemente en los mismos documentos, de alguna manera estarán relacionados entre sí. De hecho, se parte de la hipótesis de asociación (18, p.104): si un término es buen discriminante de documentos relevantes y no relevantes, un término íntimamente asociado con él también lo será. Considerando que los términos que aparecen en la consulta son buenos discriminantes de documentos relevantes y no relevantes, entonces también lo serán sus términos relacionados. Ello nos permite añadirlos a la consulta original.

Para medir el grado de asociación entre dos términos existen diferentes funciones. Las más utilizadas son las de Tanimoto (Jaccard), Coseno y Dice [18]:

$$\begin{aligned} \text{Tanimoto}(t_i, t_k) &= \frac{n_{ik}}{n_i + n_k - n_{ik}} \\ \text{Coseno}(t_i, t_k) &= \frac{n_{ik}}{\sqrt{n_i \cdot n_k}} \\ \text{Dice}(t_i, t_k) &= \frac{2 \cdot n_{ik}}{n_i + n_k} \end{aligned}$$

donde n_i y n_k son el número de documentos en los que aparece el término t_i y t_k, respectivamente, y n_{ik} el número de documentos en los que coocurren t_i y t_k. Para cualesquiera dos términos estas funciones devuelven un valor entre 0 y 1. Pues bien, hemos encontrado que, para tesauros globales los mejores resultados se obtienen con la función de Dice y añadiendo los 5 ó 10 mejores términos relacionados con toda la consulta. En este caso se obtienen mejoras en torno al 2% en precisión media y algo menos del 10% en P@10. Si la construcción del tesauro es local, los resultados son ligeramente mejores, del 5% en precisión media, y por encima del 10% en P@10.

5.2.2. Tesauros de similitud

Un tesauro de similitud es una matriz que para su construcción utiliza un mecanismo mediante el cual cada término de la colección se caracteriza por los documentos en que aparece. Es darle la vuelta al concepto clásico de los sistemas de recuperación de información. Para construir el tesauro de similitud los términos de la colección se consideran documentos, y los documentos se utilizan como términos índice. Es decir, podemos pensar que los documentos pueden servir para representar los términos. Existen varias formas para representar los términos en función de los documento y luego computar su similitud, pero en general el coste computacional es elevadísimo (tanto global como localmente). Además hemos encontrado en nuestros experimentos que se obtienen resultados prácticamente idénticos a la aplicación de tesauros de asociación.

6. CONCLUSIONES

En este trabajo se ha estudiado la actuación del usuario en nuestro sistema de recuperación, utilizando para ello el archivo de registro de las búsquedas que los usuarios han realizado. Hemos obtenido unos resultados que han sido analizados y comparados con estudios que también utilizan el análisis de archivos de registro. Todo ello considerando que DATATHÉKE es tanto un sistema de recuperación de información, en el sentido clásico de la palabra, como una biblioteca digital, en el moderno.

También se han indicado varias posibilidades para mejorar de forma automática los resultados de las consultas, basadas todas ellas en la expansión de consultas. Se han descrito las técnicas de expansión que podrían utilizarse, así como la mejoría en los resultados que se ha alcanzado utilizando la colección de pruebas proveniente de DATATHÉKE. Parece que el mecanismo computacionalmente más aceptable, que obtiene además resultados muy buenos es la pseudo realimentación de consultas, por eso es la que suele utilizarse en otros sistemas de recuperación [19].

BIBLIOGRAFÍA

[1] George W. Furnas, Thomas K. Landauer, Louis M. Gomez, y Susan T. Dumais. The vocabulary problem in human-system communication. *Comunications of the ACM*, 30(11):964-971, November 1987.

[2] E.A. Stubbs, N.E. Mangiaterra, y A.M Martínez. Internal quality audit of indexing: a new application of interindexer consistency. *Cataloguing & Classification Quaterly*, 28(4):53-70, 2000.

[3] Isidoro Gil Leiva. Consistencia en la indización de documentos entre indizadores noveles. *Anales de Documentación*, 5:99-111, 2002.

[4] Thomas A Peters. The history and development of transaction log analysis. *Library Hi Tech*, 11(2):41-66, 1993.

[5] Steve Jones, Sally Jo Cunningham, Rodger J. McNab, y Stefan Boddie. A transaction log analysis of a digital library. *International Journal on Digital Libraries*, 3(2):152-169, 2000.

[6] Bernard J. Jansen, Amanda Spink, y Tefko Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Procession & Management*, 36(2):207-227, March 2000.

[7] Dietmar Wolfram, Amanda Spink, Barnard J. Janses, y Tefko Saracevic. Vox populi: The public searching of the web. *Journal of the American Society for Information Science and Technology*, 52(12):1073-1074, 2001.

[8] Carlos G. Figuerola, José Luis A. Berrocal, y Ángel F. Zazo. Diseño de un motor de recuperación de información para uso experimental y educativo. *BID. Textos Universitaris de Bibliotecomia i Documentació*, 4, 2000. Publicación electrónica: http://www.ub.es/biblio/bid/bid04.htm [consulta: diciembre-2002].

[9] Gerald Salton. *Automatic Information Organization and Retrieval*. McGraw-Hill, New-York, 1968.

[10] Gerard Salton, A. Wong, y C.S. Yang. A vector space model for automatic indexing. *Communication of the ACM*, 18:613-620, 1975.

[11] Gerard Salton y C.S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4):351-372, 1973.

[12] Ángel F. Zazo. *Técnicas de expansión en los sistemas de recuperación de información*. PhD thesis, Departamento de Informática y Automática - Universidad de Salamanca, 2003.

[13] Amanda Spink, Dietmar Wolfram, Barnard J. Jansen, y Tefko Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226-234, 2001.

[14] Carlos G. Figuerola, Ángel F. Zazo, y José Luis A. Berrocal. La interacción con el usuario en los sistemas de recuperación de información: Realimentación por relevancia. *Scire*, 8 (1):87-94, 2002.

[15] J.J. Rocchio. Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART Retrieval System*. Experiments in Automatic Document Processing, pages 313-323. Prentice Hall, Englewoods Cliffs, N.J., 1971.

[16] Yonggang Qiu y Hans-Peter Frei. Concept-based query expansion. In R. Korfhage, E.M. Rasmussen y P. Willet (Ed*.), Proceedings of the 16th Annual International ACM-SIGIR*, pp. 160-169. ACM Press, 1993.

[17] C.J. van Rijsbergen. *Information Retrieval*. Dept. of Computer Science, University of Glasgow, second edition, 1979.

[18] Gerald Salton y M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New-York, 1983.

[19] Carol Peters, Martin Braschler, Julio Gonzalo, y Michael Kluck (Ed.). *Evaluation of Cross-Language Information Retrieval Systems. Lecture Notes in Computer Science Nº 2406*. Springer, Berlin, etc., 2002.