

El sistema de recuperación Karpanta: estudio de usuarios a través del archivo de registro

Ángel F. Zazo Rodríguez

Carlos García Figuerola

José L. Alonso Berrocal

Emilio Rodríguez

Universidad de Salamanca (España)

0.1. Resumen

Se presenta el análisis de los archivos de registro del sistema de recuperación Karpanta como un modelo del análisis de la interacción entre usuarios y sistemas en el contexto de las bibliotecas digitales.

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

0.2. Abstract

A model for the analysis of interaction between users and information retrieval systems in the context of digital libraries is proposed, based on the study of the transaction logs of the retrieval system “Karpanta”

Keywords: User studies. Information retrieval. Transaction log analysis.

1. Introducción

Los sistemas de recuperación de información se basan, en primer lugar, en conseguir una representación homogénea y procesable de documentos y consultas, y, en segundo lugar, en el cálculo de alguna función que exprese el grado de similitud entre una consulta dada y cada uno de los documentos de la colección. Aquellos documentos más similares con la consulta serán mostrados al usuario. En estos sistemas, uno de los aspectos más importantes a tener en cuenta para obtener buenos resultados es el de la formulación de consultas por parte del usuario. Ello supone especificar un conjunto de palabras o términos que expresen su necesidad informativa. Desdichadamente, la formulación de las necesidades del usuario no es un problema simple. Por un lado existen unos requerimientos del sistema para formalizar la consulta. Así, por ejemplo, la formulación de una

consulta en un sistema de recuperación booleano es muy diferente respecto de un sistema que acepte consultas en lenguaje natural. La razón está en los diferentes objetivos de diseño que se han aplicado en la construcción de cada sistema de recuperación. Por otro lado, es muy conocido el hecho de que dos personas pueden asignar diferentes palabras para referirse a los mismos conceptos.

Nuestro grupo de trabajo lleva desarrollando desde hace unos años tareas de investigación en recuperación de información, y ha puesto a disposición del público a través de Internet (<http://milano.usal.es/dtt.htm>) un motor de recuperación completamente operacional. El sistema utiliza el conocido modelo vectorial, empleando lenguaje natural para la entrada de consultas. También admite realimentación de consultas, aunque en la versión Web no está implementada esta faceta. Nuestro trabajo en este artículo se ha centrado en el análisis de los hábitos de los usuarios que utilizan nuestro sistema a través de Internet. Para ello hemos empleado el registro de las más de 45.000 consultas realizadas entre noviembre de 1999 y marzo de 2002. Se registran las búsquedas y la información relacionada, pero los usuarios permanecen anónimos.

Este artículo se organiza como sigue: primero se describe el sistema de recuperación *Karpanta*, mostrando brevemente sus fundamentos, los módulos que posee y su funcionamiento interno. En la sección 3 se indica cómo se han obtenido los datos de los archivos de registro. En la sección 4 se realiza el análisis de dichos datos. Finalmente, se indican las líneas de trabajo futuro.

2. El sistema de recuperación *Karpanta*

La recuperación de información es una disciplina que ha ganado en importancia habida cuenta del aumento de la disponibilidad de documentos en soporte electrónico y de la localización de los mismos ante una necesidad informativa dada. El estudio de los sistemas de recuperación de información es importante en los currícula de los titulados en Ciencias de la Documentación, y por tanto es aconsejable disponer de herramientas que faciliten la enseñanza de dicha materia. Este ha sido el motivo fundamental para la construcción del motor de indexación y búsqueda de información *Karpanta*, sin dejar de lado su aplicación a la investigación en este campo (Figuerola et al., 2000).

Aunque los objetivos prioritarios eran diseñar una herramienta para la docencia y la investigación, y no un motor de búsqueda operacional, el resultado ha sido lo suficientemente robusto como para ser utilizado con éxito en determinados entornos documentales. *Karpanta* es un sistema que utiliza el conocido modelo vectorial en su proceso de recuperación. En el modelo vectorial (Salton, 1968) cada documento d_i de la colección de m documentos, se representa por un vector de n elementos, $d_i = (w_{i1}, w_{i2}, w_{i3}, \dots, w_{in})$, siendo n el número de términos indi-

zables que existen en la colección documental. Cada elemento del vector indica la importancia del término correspondiente en la representación del documento. Esa importancia normalmente se computa utilizando un esquema TF-IDF (Salton y Yang, 1973). Se trata de utilizar dos ideas en cierto sentido contrapuestas. Si un término aparece muchas veces en un documento determinado (TF, *term frequency*), se convierte en un término especialmente representativo de su contenido. Sin embargo, si un término aparece en muchos documentos de la colección, tendrá poca capacidad de discriminar un documento de otro, y por tanto, es poco útil para la recuperación; esto es lo que se conoce como IDF (*inverse document frequency*). Para determinar cuáles documentos de la colección son pertinentes a la necesidad informativa del usuario es necesario comparar la consulta con cada uno de los documentos. En este sentido, el proceso aplicado para los documentos también debe aplicarse a las consultas, es decir, la consulta también se representa en el espacio vectorial de términos utilizando un vector, $q = (q_1, q_2, q_3, \dots, q_n)$, donde cada elemento q_j expresa el grado en que el término índice t_j representa las necesidades informativas de la persona que hace la consulta.

En estos sistemas, evidentemente, el resultado depende enormemente del esquema de pesado de términos y posterior cálculo de similitud que realiza el sistema. Para el pesado de términos de documentos y consultas le remitimos al estudio que se realiza en Salton y Buckley (1988). La forma más simple de obtener el grado de similitud entre una consulta y un documento es calcular el producto escalar de los vectores que los representan (van Rijsbergen, 1979). Es un proceso sencillo y uno de los más utilizados. Para que el valor de similitud se encuentre entre 0 y 1, se normalizan los vectores de documentos y consultas.

En cuanto a su arquitectura, *Karpanta* se apoya en dos módulos, uno de indización, que construye los vectores de documentos y consultas, y otro de búsqueda, que obtiene los documentos más similares a una consulta dada. La construcción de los mismos se ha realizado utilizando el SGBD Microsoft Access, por su facilidad de uso, transparencia del sistema y posibilidades docentes, a pesar del descrédito que los sistemas de bases de datos relacionales han tenido en entornos documentales (Trigueros y Higuera, 1997). La mayor parte de operaciones se implementan utilizando SQL (bajo Visual Basic), lo que posibilita modificar de manera sencilla aspectos como el cálculo de pesos o similitudes, potenciando, de nuevo, en sus ventajas docentes.

2.1. La colección documental

La base de datos contiene el vaciado 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. La mayor parte de dichas revistas están especializadas en temas relacionados

con la Biblioteconomía, la Archivística, la Informática y las Ciencias de la Documentación en general. Cada documento incluye el título, autor, descriptores, año, lengua, y, la mayoría, también el resumen del artículo en cuestión. Ahora bien, no se dispone de campos separados para cada uno de ellos, sino que toda la información se recoge en un único campo *memo* (salvo año y lengua, pues se permiten búsquedas restringidas por año o lengua del artículo). Es decir, la información es tratada como si se dispusiera en texto libre.

2.2. Módulo de indización

Karpanta posee un módulo de indización que determina cuáles son los términos índice de la colección. Este módulo realiza el preprocesado de texto, que supone la consecución de varias acciones (Fox, 1992). La primera consiste en un análisis léxico inicial con el objetivo de determinar el tratamiento que se realizará sobre números, guiones, signos de puntuación, tratamiento de mayúsculas y/o minúsculas, nombres propios, siglas, etc. El tratamiento léxico en nuestro sistema ha sido bastante sencillo, simplemente hemos sustituido las vocales acentuadas por aquellas sin acentuar, y hemos convertido todos los términos a mayúsculas. La segunda acción que se realiza es la eliminación de palabras vacías con el objetivo de reducir el número de términos índice; no se incluyen palabras que, por su poca capacidad semántica — artículos, preposiciones, conjunciones, etc. — o por su alta frecuencia, son poco significativas en el proceso de recuperación de información. El siguiente paso consiste en aplicar lematización (Hull, 1996). Se denomina lematización al proceso mediante el cual se buscan variaciones morfológicas de los términos con el objetivo de extraer la raíz común a ellos. Para nuestra colección documental hemos utilizado un *S-Stemmer* modificado ligeramente para el castellano, que unifica términos terminados en plural, pues hemos comprobado que mejora sensiblemente los resultados. El último de los pasos que se puede realizar en el preprocesado de términos es la construcción o aplicación de técnicas que permitan expandir la consulta. En nuestro caso no se ha realizado tal expansión, pues, aunque *Karpanta* posee un módulo que realimenta consultas utilizando criterios de relevancia del usuario (Figuerola et al., 2002), en la versión Web no está operativo. Finalmente, hemos considerado términos índice los restantes después de eliminar las palabras vacías y aplicar el lematizador. Una vez obtenidos los términos índice de la colección, el módulo de indización también realiza la representación de documentos y consultas utilizando el mecanismo de pesado TF-IDF con sentencias SQL, almacenando los resultados en otra tabla.

2.3. Módulo de consulta

Karpanta posee una interfaz de usuario fácil e intuitiva (figura 1). Cuando *Karpanta* recibe una consulta, ésta primero pasa por el módulo de indización, que selecciona los términos índice de la misma y asigna los pesos adecuados a los mis-



Fig. 1. Interfaz de consulta para Karpanta

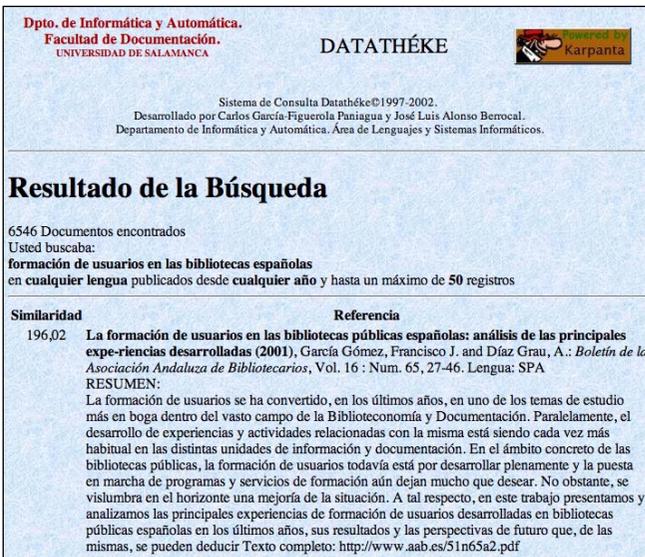


Fig. 2. Resultados de búsqueda en Karpanta

mos. Es muy importante destacar que *Karpanta* solamente acepta consultas en lenguaje natural. Es decir, el usuario plantea su necesidad informativa sin necesidad de utilizar operadores booleanos, de truncamiento, proximidad o posición.

El usuario introduce su consulta y ésta pasa por el módulo de indización. Los términos, junto con su peso, se almacenan entonces en otra tabla de la base de datos. A continuación se realiza el cálculo de la similitud entre la consulta y todos los documentos, utilizando para ello sentencias SQL. El sistema está diseñado para utilizar el producto escalar de los vectores que representan a los documentos y a la consulta, pues ello permite mostrar de forma ordenada los documentos de acuerdo al grado de similitud. En la figura 2 se aprecia el resultado de una búsqueda.

3. Archivo de registro

En *Karpanta* se recoge casi toda la actividad que realiza el usuario cuando consulta el sistema. Evidentemente, el usuario permanece totalmente anónimo: se registran sus consultas e información relacionada, pero nada más. Se han tomado los datos desde noviembre de 1999 hasta marzo de 2002. Se trata de 45.856 consultas. Para cada consulta se recoge el texto íntegro de la consulta, las opciones para año, lengua y límite de visualización de registros, la fecha y hora, y la dirección IP y, si es posible, también el nombre por dominios del ordenador con el que el usuario se conecta. Lamentablemente, el sistema no registra el grado de satisfacción del usuario respecto del resultado de la búsqueda. Sin embargo podemos estimarlo realizando el análisis de su comportamiento. Si el usuario modifica ligeramente su consulta en un intervalo de tiempo reducido —en los siguientes minutos—, será que no ha encontrado lo que buscaba, o que desea refinar más la búsqueda.

4. Análisis de los datos

Los datos del fichero de registro se han procesado automáticamente utilizando programas escritos en el lenguaje de programación Perl. Los resultados se muestran en forma de tablas o figuras.

4.1. Origen de las consultas

El primer análisis que hemos realizado trata de obtener la localización del usuario que realiza la consulta, esto es, la dirección IP o el nombre por dominios de la máquina desde la que se conecta a *Karpanta*. Examinando el acceso por dominios (tabla I), podemos observar que la mayor parte de los mismos proceden de España (.es). En un porcentaje bastante alto de casos no se ha podido determinar el dominio de procedencia (20%). Esto se debe fundamentalmente a que muchos proveedores de acceso únicamente asignan una dirección IP dinámica a sus ordenadores, y no un nombre por dominios. Lo cual indica, por otra parte, que un número alto de usuarios consultan *Karpanta* desde ordenadores que se conectan a Internet con proveedores de acceso. Considerando asimismo que la mayoría de usuarios proceden de bibliotecas y centros de documentación, podemos afirmar que muchos de ellos todavía no disponen de una conexión permanente a Internet. Dentro del dominio de España, también se muestran los resultados por subdominios (tabla II). Podemos apreciar que la mayoría de ellos son Universidades. El número más alto de consultas se realiza desde la propia Universidad de Salamanca, pero seguida muy de cerca por el dominio '.ttd.es', y en menor medida 'retevision.es' y 'uni2.es', que son proveedores de acceso a Internet. Esto nos confirma que en el periodo de tiempo analizado muchos centros documentales todavía no disponían de una conexión permanente a Internet. Señalamos también que más de 1.000 consultas se han realizado desde la Biblioteca Nacional.

| Dom. | País | Nº de consultas | % | Dom. | País | Nº de consultas | % |
|------|-------------|-----------------|-------|------|---------------|-----------------|------|
| es | España | 32.641 | 71,18 | do | R. Dominicana | 13 | 0,03 |
| IP | | 9.171 | 20,00 | gr | Grecia | 12 | 0,03 |
| net | | 1.036 | 2,26 | pa | Panamá | 12 | 0,03 |
| com | | 646 | 1,41 | be | Bélgica | 9 | 0,02 |
| mx | México | 431 | 0,94 | ie | Irlanda | 9 | 0,02 |
| ar | Argentina | 402 | 0,88 | gt | Guatemala | 8 | 0,02 |
| pt | Portugal | 390 | 0,85 | cr | Costa Rica | 6 | 0,01 |
| cu | Cuba | 151 | 0,33 | ad | Andorra | 5 | 0,01 |
| ve | Venezuela | 149 | 0,32 | bo | Bolivia | 5 | 0,01 |
| pe | Perú | 132 | 0,29 | il | Israel | 5 | 0,01 |
| co | Colombia | 121 | 0,26 | jp | Japón | 5 | 0,01 |
| br | Brasil | 75 | 0,16 | ca | Canadá | 5 | 0,01 |
| fr | Francia | 66 | 0,14 | int | | 4 | 0,01 |
| cl | Chile | 56 | 0,12 | au | Australia | 4 | 0,01 |
| uk | Reino Unido | 52 | 0,11 | ch | Suiza | 4 | 0,01 |
| it | Italia | 46 | 0,10 | no | Noruega | 4 | 0,01 |
| de | Alemania | 39 | 0,09 | lv | Letonia | 4 | 0,01 |
| org | | 38 | 0,08 | ec | Ecuador | 4 | 0,01 |
| uy | Uruguay | 26 | 0,06 | bg | Bulgaria | 3 | 0,01 |
| edu | | 19 | 0,04 | hu | Hungría | 3 | 0,01 |
| nl | Holanda | 14 | 0,03 | lu | Luxemburgo | 3 | 0,01 |

Tabla I. Origen de las consultas

| Dominio | Organismo | Nº de consultas | % |
|---------------|--------------------------------------|-----------------|-------|
| usal.es | Universidad de Salamanca | 3.862 | 11,83 |
| ttd.es | Telefónica Transmision de Datos S.A. | 3.489 | 10,69 |
| ub.es | Universidad de Barcelona | 2.060 | 6,31 |
| retevision.es | Relevision S.A. | 1.954 | 5,99 |
| ucm.es | Universidad Complutense de Madrid | 1.171 | 3,59 |
| uv.es | Universidad de Valencia | 1.109 | 3,40 |
| unex.es | Universidad de Extremadura | 1.037 | 3,18 |
| bne.es | Biblioteca Nacional de España | 1.026 | 3,14 |
| uni2.es | Lince Telecomunicaciones S.A. | 944 | 2,89 |
| um.es | Universidad de Murcia | 872 | 2,67 |
| jcyl.es | Junta de Castilla y León | 855 | 2,62 |
| us.es | Universidad de Sevilla | 792 | 2,43 |
| uc3m.es | Universidad Carlos III de Madrid | 649 | 1,99 |
| uoc.es | Universitat Oberta de Catalunya | 445 | 1,36 |
| upv.es | Universidad Politecnica de Valencia | 444 | 1,36 |
| uva.es | Universidad de Valladolid | 440 | 1,35 |
| uji.es | Universidad Jaume I | 437 | 1,34 |
| usc.es | Universidad de Santiago de Comp. | 411 | 1,26 |
| ugr.es | Universidad de Granada | 405 | 1,24 |
| unileon.es | Universidad de León | 365 | 1,12 |
| ua.es | Universidad de Alicante | 358 | 1,10 |
| ubu.es | Universidad de Burgos | 342 | 1,05 |

Tabla II. Localización de procedencia de los dominios '.es' más frecuentes

4.2. Fechas

La figura 3 muestra la distribución de consultas por fecha. El número medio de consultas por mes ha sido 1.581,24.

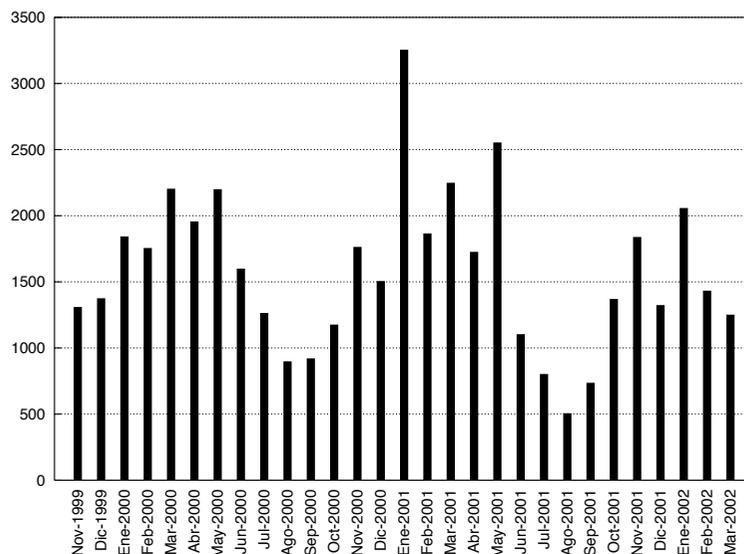


Fig. 3. Distribución de consultas por meses

4.3. Términos en las consultas

El número medio de términos por consulta es de 2,66. Es un valor muy parecido al de otros estudios sobre bibliotecas digitales (Jones et al., 2000) y sobre motores de búsqueda en Internet (Wolfram et al., 2001). En todos estos estudios se aprecia que las consultas suelen ser simples y cortas, igual que en nuestro caso. Como dato curioso, hay una consulta que posee 80 términos.

En cuanto a los términos que se utilizan en las consultas, hay un total de 10.522 términos diferentes. Para su cómputo hemos eliminado operadores booleanos y símbolos no alfanuméricos. Además se han reducido los acentos a formas no acentuadas y se han pasado todos los términos a mayúsculas. Para determinar los temas en que están más interesados nuestros usuarios, la tabla III muestra los 50 términos más frecuentes en las consultas. Resaltamos que, como era de esperar, entre los términos más frecuentes se encuentran palabras vacías —de, la, en, del, el, las, a, los, sobre—, que son eliminadas en el módulo de indexación. Es necesario destacar el efecto de la lematización en el tratamiento léxico de texto. Así, por ejemplo, los términos `biblioteca' y `bibliotecas' son tratados

internamente por el sistema como si fueran el mismo término, pues el lematizador los reduce al mismo lema. Hemos observado que algunos usuarios refinan sus consultas modificando términos en sentido de singular y plural, obteniendo evidentemente los mismos resultados (véase la sec. 4.8).

| Término | Nº de veces | Término | Nº de veces |
|---------------|-------------|----------------|-------------|
| de | 11.896 | universitarias | 851 |
| bibliotecas | 4.204 | referencia | 785 |
| la | 3.043 | del | 762 |
| informacion | 2.799 | bibliografia | 759 |
| en | 2.467 | usuarios | 757 |
| documentacion | 1.994 | datos | 739 |
| biblioteca | 1.741 | fuentes | 689 |
| internet | 1.400 | recuperacion | 652 |
| archivos | 1.112 | traduccion | 635 |
| gestion | 857 | historia | 618 |

Tabla III. Términos más frecuentes en las consultas

4.4. Consultas mal formuladas

Ya hemos indicado que *Karpanta* acepta consultas en lenguaje natural, de modo que cualquier tipo de operador booleano, de truncamiento, posición o proximidad no es entendido por el sistema. De hecho *Karpanta* trata todas las palabras como términos de la consulta, y no tiene conocimiento de tales operadores. Tampoco tiene en cuenta el orden en que han sido introducidos los términos de la búsqueda. En varias de las consultas hemos detectado operadores simbólicos (+, &, l, etc.); aunque, cuando se realiza el procesado de texto de la consulta, el módulo de indización elimina todo carácter no alfanumérico, y por tanto son eliminados. El número de consultas que utilizan algún tipo de operador como los indicados es de 4.016 (8,76 % sobre el total de consultas). Consideramos que es un número elevado de consultas a pesar de que el interfaz de *Karpanta* (figura 1) indica expresamente que la consulta debe introducirse en lenguaje natural. Los

| Operadores | Nº de veces | % |
|-------------------|-------------|------|
| AND, and, +, & | 3.835 | 8,36 |
| OR, or, l | 235 | 0,51 |
| Quotation (" , ') | 50 | 0,11 |
| NOT, not | 18 | 0,04 |
| Truncamiento (*) | 10 | 0,02 |
| Selección (IN) | 6 | 0,01 |
| Proximidad (NEAR) | 4 | 0,01 |
| Algún operador | 4.016 | 8,76 |

Tabla IV. Operadores booleanos, de truncamiento, posición, selección o proximidad

usuarios que diseñan este tipo de consultas esperan un resultado que el sistema *Karpanta* no puede satisfacer. En la tabla IV se indica el número de consultas en las que aparecen esta clase de operadores. No se ha contabilizado el operador simbólico - (correspondiente en muchos casos al operador NOT) porque hemos observado que no se utiliza. 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.

4.5. Opciones por defecto

Hemos observado que la mayoría de usuarios acepta las opciones por defecto que presenta el interfaz de consulta, esto es, visualizar 50 registros, sin limitar por año y sin buscar por una determinada lengua. El número total de consultas que no han modificado estas opciones ha sido 22.512 (49,09 %). Es decir, algo más de la mitad de usuarios modifica las opciones por defecto, lo cual nos sugiere que nuestros usuarios están bien formados, y saben lo que quieren buscar y cómo encontrarlo. La tabla V indica el número de situaciones que se han dado, y su porcentaje respecto del total de consultas.

| Opción | Nº de consultas | % |
|---|-----------------|-------|
| Opciones por defecto | 22.512 | 49,09 |
| Visualización diferente de 50 registros | 8.017 | 17,48 |
| Limitación por lengua | 16.869 | 36,79 |
| Limitación por año | 10.204 | 22,25 |
| Visualización diferente y limitación por lengua | 4.500 | 9,81 |
| Visualización diferente y limitación por año | 2.853 | 6,22 |
| Limitación por lengua y año | 6.285 | 13,71 |
| Ningún parámetro por defecto | 1.892 | 4,13 |

Tabla V. Número de consultas con relación a las opciones por defecto

4.6. Usuarios

En la figura 4 se indica el número de usuarios durante los meses analizados. Resaltamos que se consideran usuarios diferentes si la dirección IP del ordenador desde donde se conectan es diferente. Observamos una evolución similar a la distribución del número de consultas por meses. El número total de usuarios, es decir, de direcciones IP diferentes, ha sido 7.223. El número medio de usuarios por mes ha sido 359,45. Es muy interesante determinar el índice de asiduidad o visitas repetidas por usuario. Se trata de mostrar cuántas consultas se han realizado desde el mismo ordenador, en el periodo de tiempo analizado, para así obtener el grado de aceptación de nuestro servicio. La tabla VI muestra los resultados. Podemos observar que un tercio de los usuarios solamente se ha conectado una vez a *Karpanta*. Hay que resaltar de nuevo que los datos se han calculado tomando la dirección IP del ordenador desde donde se conecta el usuario; sin embargo, muchos usuarios se conectan con proveedores de acceso que asignan direcciones IP dinámicas, por lo que ese valor puede llegar a ser engañoso. Un dato importante es que el 10,41 % de usuarios ha realizado más de 10 consultas desde la misma localización. Esto nos indica una buena aceptación de nuestro servicio.

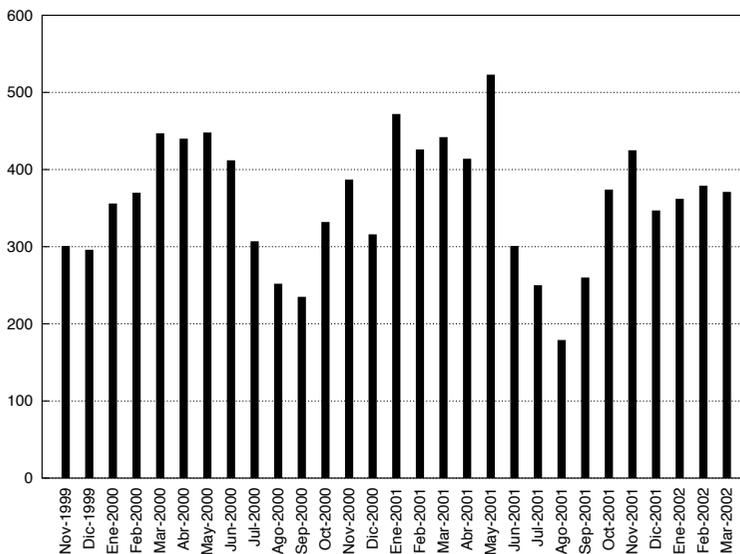


Fig. 4. Número de usuarios diferentes cada mes

| Frecuencia de consultas desde la misma dirección IP (usuario) | Nº de usuarios | % |
|---|----------------|-------|
| 1 | 2.521 | 34,90 |
| 2 | 1.488 | 20,60 |
| 3 | 805 | 11,14 |
| 4 | 535 | 7,41 |
| 5 | 329 | 4,55 |
| 6 | 273 | 3,78 |
| 7 | 175 | 2,42 |
| 8 | 144 | 1,99 |
| 9 | 121 | 1,68 |
| 10 | 80 | 1,11 |
| +10 | 752 | 10,41 |

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

4.7. Sesiones

Muchos usuarios realizan una única consulta por sesión. 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. Una sesión se inicia a una determinada hora, y acaba en otra: la duración la marca el comportamiento del usuario. La elección de la duración de una sesión debe tener en cuenta el tiempo necesario

para que el usuario pueda visualizar los registros devueltos, imprimirlos si fuera necesario, o realizar algún otro tipo de acción relacionada. Nosotros hemos considerado que normalmente estas operaciones no deben suponer una duración más larga de 30 minutos para una consulta particular. Es decir, si un usuario realiza una consulta en un momento dado, y realiza otra consulta dentro de los 30 minutos siguientes, se trata de una consulta

en la misma sesión; si realiza la consulta después de 30 minutos, se consideran dos sesiones diferentes. En la tabla VII se indica el número de consultas que se han realizado por sesión. Observamos que el 39,68 % de los usuarios solamente hace una consulta por sesión. Este resultado era de esperar, si consideramos que el 34,90 % de los usuarios solamente se ha conectado una vez a *Karpanta*. La media de consultas por sesión es de 2,89. Curiosamente, hay una sesión de 156 consultas. Cuando el usuario realiza más de una consulta por sesión puede ser porque desea refinar más en su búsqueda, modificando los términos de la misma. Este aspecto lo trataremos en la sección siguiente.

4.8. Refinamiento de consultas

Cuando el usuario realiza más de una consulta por sesión (tabla VII), es importante analizar su comportamiento en lo que se refiere a la reformulación de la consulta original para obtener mejores resultados. En este sentido, aplicamos las mismas restricciones que al hablar de sesiones: consideramos que la sucesión de consultas en una misma sesión se realiza dentro de los 30 minutos siguientes. Para no enmascarar resultados, hemos eliminado expresamente aquellas consultas iguales a la consulta anterior, ya que, curiosamente, se ha detectado que existe un porcentaje muy alto de usuarios que repiten exactamente la misma consulta (se trata de 8.265 consultas, es decir, el 18,00 % sobre el total de consultas).

La parte izquierda de la tabla VIII muestra el número de términos comunes en consultas consecutivas de la misma sesión. El 59,57 % de las consultas realizadas en una misma sesión no contiene ningún término común, es decir, se trata de consultas completamente diferentes. El resto de consultas sucesivas tiene al menos un término común, lo cual supone el 40,43 %. Esto está de acuerdo con lo observado en los sistemas de recuperación de información tradicionales, en los que el nivel de reformulación de consultas es también elevado.

| Nº de consultas por sesión | Frecuencia | % |
|----------------------------|------------|-------|
| 1 | 6.301 | 39,68 |
| 2 | 3.729 | 23,49 |
| 3 | 2.046 | 12,89 |
| 4 | 1.255 | 7,90 |
| 5 | 784 | 4,94 |
| 6 | 527 | 3,32 |
| 7 | 345 | 2,17 |
| 8 | 234 | 1,47 |
| 9 | 161 | 1,01 |
| 10 | 105 | 0,66 |
| +10 | 391 | 2,46 |

Tabla VII. Número de consultas por sesión

Visto que la reformulación de consultas es una practica habitual, y que el número medio de términos por consulta es de 2,66, parece claro que los usuarios refinan sus necesidades informativas con cambios muy pequeños: añadiendo algún término, modificando algún otro, quitando otro, etc. Este extremo se puede comprobar en la tabla IX, en la que se muestra el número de términos que se modifican en consultas sucesivas. Se han tomado exclusivamente las consultas que tienen algún término común. Un 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, y un valor de -1 indica que se ha eliminado un término de la consulta. Podemos observar que en la reformulación de consultas predomina la sustitución de un término por otro.

5. Líneas futuras de investigación

En este trabajo se ha estudiado la actuación del usuario en nuestro sistema de recuperación *Karpanta*, utilizando para ello el archivo de registro de las consultas que los usuarios lanzan al sistema. En cuanto a las líneas de trabajo futuro, destacamos dos. La primera de ellas consiste en el análisis del comportamiento del usuario en relación con los documentos que seleccione para visualizar. Actualmente el sistema muestra toda la información del documento al usuario. Sería conveniente que se mostrase solamente el título del documento, y se le obligara a seguir un enlace para visualizar el documento completo. Esto nos permitiría analizar aspectos más interesantes del comportamiento del usuario. La segunda trata de investigar el comportamiento del usuario en relación con la realimentación de consultas. En este sentido, se pueden diseñar dos experimentos diferentes. Uno es la utilización de un botón similar al de muchos buscadores (“páginas similares”, “*more like this*”, etc.). Ese aspecto está íntimamente relacionado con la clasificación y *clustering* de documentos. El otro es la aplicación de realimentación de consultas con crite-

| Nº de términos comunes | % |
|------------------------|-------|
| 0 | 59,57 |
| 1 | 17,08 |
| 2 | 10,34 |
| 3 | 5,18 |
| 4 | 2,64 |
| 5 | 1,20 |
| 6 | 0,74 |
| 7 | 0,84 |
| 8 | 0,64 |
| 9 | 0,47 |
| 10 | 0,29 |
| +10 | 1,02 |

Tabla VIII. Número de términos comunes en consultas sucesivas

| Nº de términos añadidos | % |
|-------------------------|-------|
| +5 | 1,12 |
| 5 | 0,84 |
| 4 | 1,81 |
| 3 | 4,16 |
| 2 | 9,30 |
| 1 | 21,38 |
| 0 | 38,97 |
| -1 | 11,57 |
| -2 | 5,83 |
| -3 | 2,39 |
| -4 | 1,29 |
| -5 | 0,63 |
| +(-5) | 0,80 |

Tabla IX. Número de términos añadidos en consultas sucesivas

rios de relevancia del usuario. Para ello, el interfaz de consulta debería disponer de sendas casillas de verificación para que el usuario pudiera marcar documentos relevantes y no relevantes, y poderse, así, observar su comportamiento.

6. Referencias

- Figuerola, C. G.; Berrocal, J. L. A. ; Zazo Rodríguez, A. F. (2000). Diseño de un motor de recuperación de información para uso experimental y educativo. // *BID: Textos universitarios de bibliotecología i documentació*. 4 (2000). URL: <<http://www.ub.es/biblio/bid/bid04.htm>>.
- Figuerola, C. G.; Zazo Rodríguez, A. F. ; Berrocal, J. L. A. (2002). La interacción con el usuario en los sistemas de recuperación de información: realimentación por relevancia. // *Scire*. 8:1 (en.-jun. 2002) 87-94.
- Fox, C. (1992). Lexical analysis and stoplist. // Frakes, W.B., Baeza-Yates, R. (editors). *Information retrieval: Data Structures and Algorithms*. Englewood Cliffs (NJ): Prentice-Hall Inc., 1992. 102-130.
- Hull, D. A. (1996). Stemming algorithms: A case study for detailed evaluation. // *Journal of the American Society for Information Science*. 47:1 (1996) 70-84.
- Jansen, B. J.; Spink, A.; Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. // *Information Processing & Management*. 36:2 (2000) 207-227.
- Jones, S.; Cunningham, S. J. ; McNab, R. ; Boddie, S. (2000). A transaction log analysis of a digital library. // *International Journal of Digital Library*. 3 (2000) 152-169.
- Peters, T. A. (1993). The history and development of transaction log analysis. // *Library Hi Tech*. 11:2 (1993) 41-66.
- Salton, G. (1968). *Automatic Information Organization and Retrieval*. New-York: McGraw-Hill, 1968.
- Salton, G. ; Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. // *Information Processing & Management*. 24:5 (1988) 513-523.
- Salton, G.; Yang, C. S. (1973). On the specification of term values in automatic indexing. // *Journal of Documentation*. 29:4 (1973) 351-372.
- Spink, A.; Wolfram, D. ; Jansen, B. J. ; Saracevic, T. (2001). Searching the web: The public and their queries. // *Journal of the American Society for Information Science and Technology*. 52:3 (2001) 226-234.
- Trigueros, J.; Higuera, R. (1997). Bases de datos relacionales versus bases de datos documentales. // *Boletín de la Asociación Andaluza de Bibliotecarios*. 13:49 (1997) 43-57.
- van Rijsbergen, C. (1979). *Information Retrieval*, 2nd edition. Glasgow: University of Glasgow, Dept. of Computer Science, 1979. También en línea, URL: <<http://www.dcs.gla.ac.uk/Keith/>>.
- Wolfram, D. ; Spink, A. ; Jansen, B. J. ; Saracevic, T. (2001). Vox populi: The public searching of the web. // *Journal of the American Society for Information Science and Technology*. 52:12 (2001) 1073-1074.