

Sistema interactivo de búsqueda de respuestas con sugerencia de términos

Angel Zazo, José Luis Alonso Berrocal, Carlos G. Figuerola, y Raquel
Gómez Díaz

Grupo de investigación REINA - Universidad de Salamanca
C/ Francisco de Vitoria, 6-16, 37008 – Salamanca
{zazo, berrocal, figue, rgomez}@usal.es

Resumen Este trabajo analiza el comportamiento del usuario en un sistema interactivo de búsqueda de respuestas que incluye mecanismos de sugerencia de términos. El análisis se lleva a cabo utilizando metodologías de evaluación aplicadas en experimentos de este tipo. Se organizó un grupo de 16 usuarios, cada uno de ellos debía encontrar la respuesta a ocho preguntas, más dos de entrenamiento; sólo en la mitad se aplicaba el mecanismo de sugerencia de términos. Los resultados muestran que la posibilidad de añadir términos relacionados con los de la pregunta fue bien valorada por los usuarios, que la consideraron útil, y con la que se obtuvieron respuestas más precisas. Hasta el momento hemos utilizado relaciones de coocurrencia de términos para crear un tesoro de términos relacionados con los de la pregunta, que se muestran al usuario mediante una capa flotante para, si lo desea, añadirlos a su pregunta, pero estamos trabajando también en técnicas de extracción de información.

1. Introducción

Una de las tareas más avanzadas en Recuperación de Información es aquella que busca respuestas precisas a necesidades de información concretas. Esto es lo que se conoce como búsqueda de respuestas (BR), *question answering* en inglés. Los sistemas de BR deben encontrar un fragmento de texto mínimo que responda a la necesidad informativa concreta. Por eso, algunas aproximaciones a este problema están basadas en la recuperación de pasajes o fragmentos de texto [12,13]. La respuesta puede ser extraída o generada después mediante algún proceso automático o, si el sistema no es capaz de proporcionar una respuesta válida de una manera totalmente autónoma, debe disponerse de algún proceso de interacción con el usuario para que este pueda satisfacer su necesidad informativa.

Muchos usuarios utilizan sistemas estándar de recuperación de información, como los motores de búsqueda de Internet, para los procesos de búsqueda de respuestas. Es un escenario bastante frecuente. La gran cantidad de documentos y la habilidad del usuario para refinar sus búsquedas permite a estos reconocer la respuesta buscada dentro de los documentos recuperados. Dada la complejidad de este tipo de búsquedas, el nivel de participación del usuario, en términos de cantidad y de calidad, es mayor que para otros sistemas de recuperación. La

definición de una acertada estrategia de búsqueda y su formulación requieren un nivel de experiencia mayor. Sin ningún tipo de asistencia, debe ser el propio usuario el que, en su proceso de refinamiento, dé solución a todos los problemas que va encontrando para plasmar adecuadamente su necesidad informativa.

El objetivo de este trabajo ha sido aplicar tecnologías de análisis de textos y mecanismos de presentación de información, como elementos de ayuda para que los usuarios sean capaces de plasmar de forma más precisa su necesidad informativa. Partiendo de los términos de búsqueda, o al menos de los caracteres que va tecleando el usuario, han debido resolverse varios aspectos, como qué mecanismos de sugerencia de términos pueden aplicarse, cuáles pueden ser los términos sugeridos, cómo presentar al usuario la información que pueda ayudarle en el planteamiento de su consulta, o cómo realizar la presentación de resultados.

Uno de los objetivos concretos ha sido determinar si la sugerencia de términos ayuda al usuario a encontrar mejores respuestas a las preguntas formuladas, en cuanto a precisión y rapidez. Ahora bien, la experimentación con usuarios reales entraña algunos inconvenientes. Por ejemplo, una misma pregunta no puede ser formulada al mismo usuario más que una sola vez, por el efecto de aprendizaje que supone haberla contestado previamente. Para evitar efectos adversos lo normal es utilizar un orden pseudo-aleatorio de preguntas y sistemas a evaluar. En la sección 3 volveremos sobre este aspecto.

En la Sección 2 se describen los módulos de sugerencia de términos. En la Sección 3 se describe el diseño de nuestros experimentos. En la Sección 4 analizamos los resultados, y en la 5 se presentan las conclusiones y comentarios.

2. Sugerencia de términos

Existen varios mecanismos para asistir al usuario en el planteamiento de sus necesidades informativas, en general mediante la sugerencia o exploración de términos [9], unas veces de construcción manual, como los tesauros de los distribuidores comerciales de bases de datos, otras de construcción automática, como la utilización de jerarquías de términos extraídas de agrupación documental, relaciones de co-ocurrencia entre términos, etc. En los actuales buscadores de Internet el usuario puede verse ayudado principalmente de dos maneras: mediante el *autocompletado* de términos o frases según se van introduciendo caracteres en el campo de búsqueda (Google, Yahoo!), o mediante la *sugerencia de términos relacionados* con los que ha introducido en su búsqueda (lo ofrecen Google y Yahoo! en botones especiales). Esos han sido también los módulos que hemos desarrollado en nuestro trabajo.

Nuestro *módulo de autocompletado* se lanzaba automáticamente al ir tecleando el usuario los términos de búsqueda. Se implementó una capa flotante debajo del campo de búsqueda en la que aparecían términos que comenzaban por los caracteres tecleados hasta ese momento. Se permitía incluso el autocompletado de varios términos. Sin embargo, una vez desarrollado dicho módulo, decidimos prescindir de él, pues nos dimos cuenta de que en realidad no suponía ninguna ventaja para los usuarios: las preguntas estaban formuladas de antemano, con

lo que el usuario ya sabía los términos que tenía que buscar. Por otra parte, entre los términos sugeridos aparecían con demasiada frecuencia términos que claramente eran errores de tecleo, y eso podía llegar a confundir al usuario en vez de ayudarlo. Este fue otro motivo para no incluirlo en la interfaz final.

El *módulo de sugerencia de términos relacionados* con los de la consulta se presentaba igualmente como una capa flotante debajo del campo de búsqueda, y contenía términos que podían ser añadidos a la búsqueda. Decidimos que no apareciera de manera automática, sino que se lanzase cuando el usuario tuviese la necesidad de que el sistema le sugiriese términos. Utilizamos para ello el carácter dos puntos (:) en vez de botones especiales. Para el desarrollo de este módulo utilizamos relaciones de co-ocurrencia: dos términos están relacionados si aparecen en los mismos contextos. Ello nos permite crear un tesoro de términos relacionados con los de la consulta. Para medir el grado de relación entre dos términos existen varias funciones [10]; hemos utilizado la de Dice porque en trabajos previos hemos comprobado su buen comportamiento [15]. Como unidad de co-ocurrencia empleamos pasajes de texto: justamente los términos sugeridos proceden de los primeros pasajes recuperados teniendo en cuenta los términos introducidos por el usuario antes de escribir el carácter dos puntos.

La ecuación (1) muestra la función de asociación de Dice entre un término t_q de la consulta y un término t_k con el que co-ocurre en uno o más pasajes; n_q y n_k son el número de pasajes en los que aparecen t_q y t_k , respectivamente, y n_{qk} el número de pasajes en los que co-ocurren.

$$\text{Dice}(t_q, t_k) = \frac{2 \cdot n_{qk}}{n_q + n_k} \quad (1)$$

Para cada término de los pasajes se calculó el valor de asociación con toda la consulta sumando los valores individuales de asociación con cada uno de los términos de la misma. Finalmente, los términos con valores más altos de asociación se mostraban al usuario (véase la Fig. 1). Comprobamos, no obstante, que los términos más frecuentes en los pasajes introducían un efecto negativo en la sugerencia de términos. Los términos más frecuentes son también los candidatos a ser los mejor relacionados con el resto de términos, y por tanto debieran ocupar posiciones destacadas al ser sugeridos al usuario, pero el usuario los encontraría tan generales que le servirían de poco para refinar su consulta. En la ecuación (2) se evita el efecto de valores altos de IDF [11], siendo N el número de pasajes considerados a la hora de obtener las relaciones de co-ocurrencia; su valor se fijó en 30, igual al número de pasajes que se mostraban al usuario en la interfaz de búsqueda.

$$\text{Dice}^{\text{IDF}}(t_q, t_k) = \frac{2 \cdot n_{qk}}{n_q + n_k} \cdot \log \frac{N}{n_k} \quad (2)$$

3. Experimentos

En nuestro experimento hemos seguido la metodología utilizada en varias tareas de las conferencias TREC y CLEF para la evaluación de sistemas interactivos

de búsqueda de respuestas [3, 5]. Hemos utilizado la colección de documentos en español procedente de la Agencia EFE de noticias correspondiente a los años 1994 y 1995, utilizada con fines de investigación bajo licencia de ELRA y utilizada extensamente en las conferencias CLEF [1]. Se trata de 454.045 documentos etiquetados bajo SGML con varios campos, como número de documento, título, texto de la noticia, fecha, categoría, etc. En nuestros experimentos solamente hemos utilizado los campos de título, texto y fecha. Tal como se ha indicado en la introducción, la búsqueda de respuestas muy a menudo utiliza pasajes de texto como unidades de búsqueda. Es el enfoque que nosotros hemos aplicado, entre otras cosas porque para un usuario es más sencillo localizar la respuesta a una pregunta en un pasaje de texto que tener que leer el documento completo.

En varios experimentos interactivos se ha comprobado que la posibilidad de ver el documento completo del que deriva el pasaje debiera ser una de las características deseables del sistema, sin duda, porque al dividir el documento en pasajes se pierde un contexto fundamental que sirve para verificar que la respuesta es correcta [2, 7]. En nuestro caso hemos decidido intencionadamente no disponer de esa posibilidad, y centrarnos exclusivamente en la sugerencia de términos.

Cada pasaje de texto estaba formado por al menos un párrafo. Después de varias pruebas se decidió que cada pasaje debía tener al menos 60 palabras. Si al ir dividiendo el documento en párrafos, alguno de los pasajes tenía menor número de palabras, se le iba concatenando otro párrafo, de manera que el pasaje tuviera al menos 60 palabras. El número total de pasajes fue de casi dos millones, con una media de 77,8 palabras por pasaje.

En cuanto a las preguntas del experimento, en la Tabla 1 pueden verse las que se utilizaron. Procedían de las tareas interactivas de las conferencias CLEF de los años 2004 y 2005, aunque aquí el entorno ha sido monolingüe. En esas tareas el número de preguntas fue justamente el doble, convirtiéndose en un experimento largo y pesado para los usuarios [14], por eso seleccionamos un número de preguntas menor. Las preguntas 9 y 10 se emplearon como preguntas de entrenamiento para que los usuarios se familiarizasen con la interfaz de consulta. Todas las preguntas tenían respuesta en la colección documental.

Para poder dar por correcta una respuesta el usuario debía indicar el pasaje de texto en el que la había encontrado. Si no indicaba el pasaje, o este no respondía verdaderamente a la pregunta, se consideró que la respuesta no era válida en sentido estricto, salvo que el usuario ya conociese la respuesta de antemano. Para asegurar este punto, se pasó previamente un cuestionario en el que el usuario debía indicar su conocimiento sobre cada pregunta.

3.1. Sistema de recuperación

No hemos utilizado un sistema de búsqueda de respuestas en nuestros experimentos, sino que hemos adaptado un sistema clásico de recuperación de información para esta tarea. Varios experimentos como el nuestro se han desarrollado con sistemas genéricos de búsqueda [6], de hecho, es muy frecuente que los usuarios de Internet utilicen buscadores generales para la búsqueda de respuestas. Es de

Tabla 1. Preguntas del experimento.

Nº	Pregunta
1	¿Qué institución inició la campaña europea de la juventud contra el racismo?
2	¿Qué iglesia ordenó mujeres sacerdote en marzo de 1994?
3	¿Quién era primer ministro de Noruega cuando se celebró el referéndum sobre su posible incorporación a la UE?
4	¿Cuándo se estima que ocurrió el Big Bang?
5	¿Cuánto costó el Túnel del Canal?
6	¿Quién es el director gerente del Fondo Monetario Internacional?
7	Fecha de la muerte de Lenin
8	¿Quién ganó el Premio Nobel de Literatura en 1994?
9	¿Cuándo abdicó Eduardo VIII?
10	¿Con el nombre de qué enfermedad se corresponde el acrónimo BSE?

señalar que en algunos casos la utilización de filtros propios de la búsqueda de respuestas para seleccionar el tipo de respuesta buscada (nombres, referencias temporales o cantidades) no han conseguido mejorar los resultados en procesos interactivos [7].

En cuanto al proceso de indización, se eliminaron acentos, se redujeron las palabras a minúsculas, y se consideraron vacías las palabras que aparecían en más del 5% de los documentos; no se aplicó lematización, aunque se podía utilizar el asterisco como operador de truncamiento por la derecha, pero fue poco utilizado.

3.2. El experimento interactivo

La interacción con el usuario presenta aspectos importantes que deben ser destacados en nuestros experimentos. Por un lado, se trata de un experimento de búsqueda de respuestas, y en este caso, el conocimiento previo de su temática o de su posible respuesta condiciona la manera en la que los usuarios plantean sus búsquedas, de ahí el cuestionario previo. Por otro lado, para dar solución a cada una de las preguntas hay que proporcionar un tiempo suficiente, pero finito; de otra manera, el usuario puede dedicar demasiado tiempo a las primeras preguntas y llegar muy cansado a las últimas. Para resolver cada una de las preguntas se contó con un tiempo máximo de cinco minutos.

La mayoría de experimentos interactivos presentan dos sistemas para ser evaluados; ambos son idénticos salvo en alguna de las funcionalidades que uno de ellos presenta y el otro no. Esa funcionalidad es la que se desea evaluar. En nuestro caso ha sido la posibilidad de disponer de sugerencia de términos relacionados con los que introduce el usuario. Hemos denominado **sistema A** al sistema que no permite la sugerencia de términos, y **sistema B** al que sí la permite. Cada usuario debía resolver la mitad de las preguntas en un sistema y la otra mitad en el otro. Todos los usuarios eran instruidos previamente, disponiéndose de dos preguntas de entrenamiento, una en cada sistema. Para evitar el efecto que el orden de las preguntas o del sistema pudiera tener en el desarrollo del experimento, se utilizó un orden pseudo-aleatorio de actuación para cada uno de los usuarios,

Tabla 2. Esquema de actuación de cada usuario.

Nº	Entren.	Experimento	Nº	Entren.	Experimento
u1	B:10 A:9	B:4-7-5-8 A:1-3-2-6	u9	A:10 B:9	A:4-7-5-8 B:1-3-2-6
u2	A:9 B:10	A:3-5-7-1 B:8-4-6-2	u10	B:9 A:10	B:3-5-7-1 A:8-4-6-2
u3	A:10 B:9	A:1-3-4-6 B:2-8-7-5	u11	B:10 A:9	B:1-3-4-6 A:2-8-7-5
u4	A:9 B:10	A:5-2-6-3 B:4-7-1-8	u12	B:9 A:10	B:5-2-6-3 A:4-7-1-8
u5	B:10 A:9	B:7-6-2-4 A:3-5-8-1	u13	A:10 B:9	A:7-6-2-4 B:3-5-8-1
u6	B:9 A:10	B:8-4-3-2 A:6-1-5-7	u14	A:9 B:10	A:8-4-3-2 B:6-1-5-7
u7	A:10 B:9	A:6-1-8-7 B:5-2-4-3	u15	B:10 A:9	B:6-1-8-7 A:5-2-4-3
u8	B:9 A:10	B:2-8-1-5 A:7-6-3-4	u16	A:9 B:10	A:2-8-1-5 B:7-6-3-4

como puede verse en la Tabla 2. Este modelo de actuación ha sido aplicado en experimentos similares en tareas interactivas de TREC y CLEF [3, 5].

El **sistema A** consistió en presentar al usuario 30 pasajes de texto en los que posiblemente se encontrase la respuesta a la pregunta. Se eligió este número para que fuese mayor que dos pantallas de resultados de motores de Internet (20 documentos), número máximo de pantallas que los usuarios suelen ver. El **sistema B** era idéntico al *sistema A*, pero incluía el módulo de sugerencia de términos: el usuario escribía los términos de la consulta que le parecían convenientes y luego podía, si así lo deseaba, escribir el carácter dos puntos para que el sistema ofreciese términos relacionados (véase la Fig. 1).

En experimentos interactivos es importante registrar las impresiones de los usuarios, y para ello hemos utilizado tres cuestionarios: uno previo al experimento, otro al finalizar la sesión de preguntas con cada sistema, y uno más al final. El primero tenía el objetivo de determinar el grado de experiencia de cada usuario en sistemas de búsqueda de información y en el conocimiento previo de las preguntas. En el cuestionario posterior a cada sistema se realizaban preguntas sobre dicho sistema, haciendo énfasis en la funcionalidad que se quería resaltar. El cuestionario final tenía que ver con todo el experimento, una comparativa de ambos sistemas y comentarios de cada uno de ellos. Más adelante se describen los resultados de dichos cuestionarios.

Para nuestros experimentos formamos un grupo de 16 usuarios, con una edad media de 21,9 años, todos ellos con amplia experiencia en sistemas de búsquedas en OPAC y en Internet, es decir, sistemas que utilizan los términos de búsqueda como términos índice. Se trataba de alumnos del *Grado en Información y Documentación* y de la *Licenciatura en Documentación* de la Universidad de Salamanca, lo cual ha dotado a este estudio de un valor añadido. En la Tabla 3 se pueden ver los resultados del cuestionario previo al experimento. Se indican valores medios en una escala de 1 a 5.

El bagaje cultural de los usuarios puede llevar a distorsiones en los resultados de un experimento interactivo de búsqueda de respuestas, debido a la propia temática de las preguntas. En este sentido, el cuestionario previo sobre la familiaridad del usuario con las preguntas mostró que para el 65% de las mismas los usuarios no tenían idea de cuáles podrían ser sus respuestas. Para el 34% de las preguntas los usuarios indicaron que el tema les sonaba, pero que no conocían

Tabla 3. Resultados del cuestionario previo al experimento.

Número de años que realiza búsquedas en línea	8,2
Experiencia de búsquedas en OPAC	4,0
Experiencia en buscadores de Internet	4,7
Frecuencia de búsqueda	4,8
Satisfacción en la búsqueda de información	4,0

la respuesta. Solamente dos usuarios dijeron saber la respuesta de una pregunta (la pregunta 2). En resumen, salvo en estos dos casos, para el resto de preguntas no era conocida su respuesta exacta.

3.3. La interfaz de búsqueda

Para el diseño de la interfaz utilizamos los típicos formularios web, un entorno muy conocido por los usuarios. La Figura 1 muestra la interfaz interactiva de búsqueda para el *sistema B* con la pregunta de entrenamiento número 10 justamente después de que el usuario haya añadido a sus términos de búsqueda “enfermedad” y “BSE” los dos primeros términos sugeridos, “locas” y “vacas”. La interfaz permitía modificar la búsqueda en cualquier momento dentro del tiempo restante, mostrado en la parte superior derecha. En la parte central aparecían

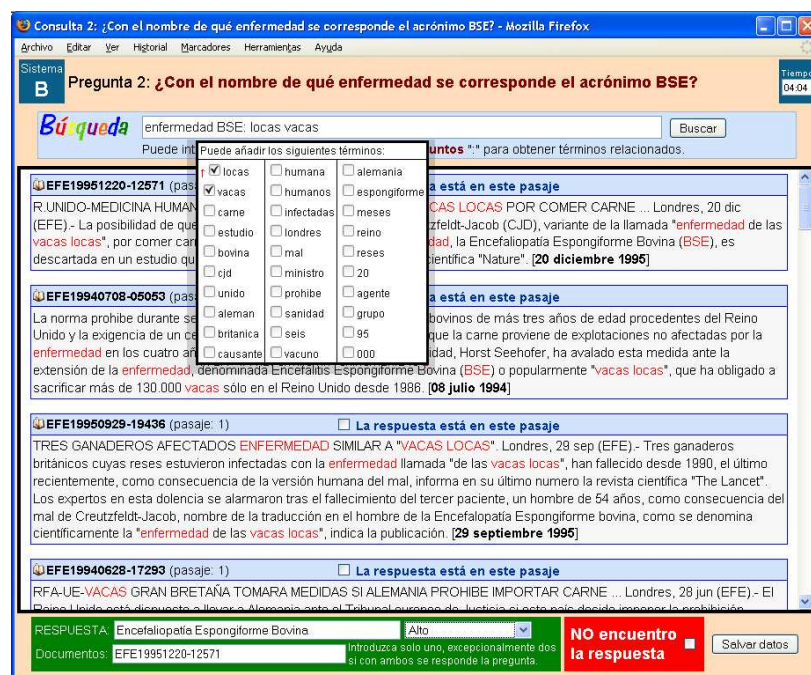


Figura 1. Interfaz de consulta.

los pasajes devueltos por el sistema de recuperación. Un aspecto destacado de la interfaz de usuario era que mostraba resaltados en color rojo aquellos términos no vacíos de la pregunta que aparecían en los pasajes de texto. De esta manera el usuario podía fácilmente localizar la respuesta a las preguntas. La interfaz de búsqueda del *sistema A* era idéntica a la del *sistema B*, solamente que no se permitía la sugerencia de términos.

Cuando el usuario encontraba la respuesta a la pregunta debía marcar el pasaje correspondiente y escribir la respuesta en la parte inferior de la ventana, o en caso de darse por vencido, marcar que no encontró la respuesta, y seguir con una nueva pregunta. De cualquier modo, una vez superados los cinco minutos el sistema automáticamente mostraba una última posibilidad de introducir una respuesta o marcar que no se había encontrado.

3.4. Juicios

Decidimos utilizar los mismos juicios de valoración de las respuestas que los utilizados en la búsqueda de respuestas [8]: una respuesta es correcta (R) si es exacta y el documento indicado la respalda; una respuesta es inexacta (X) si el documento contiene la respuesta, pero ésta no es correcta por incompleta o por contener más texto del necesario; una respuesta sin respaldo (U) es aquella que es correcta pero sin respaldo del documento. El resto son incorrectas.

Considerando esta división se tienen dos tipos de medidas de precisión: estricta y relajada. La precisión en sentido estricto considera únicamente como correctas las respuestas exactas (R), mientras que en sentido relajado también considera correctas las que no tienen respaldo (R+U). Por el diseño de nuestra interfaz sólo se dieron respuestas sin respaldo en dos ocasiones, pues se obligaba al usuario a incluir el pasaje al dar una respuesta. Por este motivo, en los resultados utilizaremos la precisión en sentido relajado.

4. Resultados

4.1. Precisión

En la Tabla 4 se aprecia que el *sistema B* es superior en un 23% al *sistema A*. Ahora bien, no todos los usuarios utilizaron la sugerencia de términos en el *sistema B*. El número de preguntas para las que se utilizó la sugerencia de términos fue de 41, de un total de 64, y fueron correctas en sentido relajado 36, el 87,8%. De las 87 preguntas restantes que no utilizaron la sugerencia de términos

Tabla 4. Precisión en sentido relajado y tiempo medio por consulta.

Sistema	Precisión	Tiempo
A	0,61	65
B	0,75	72

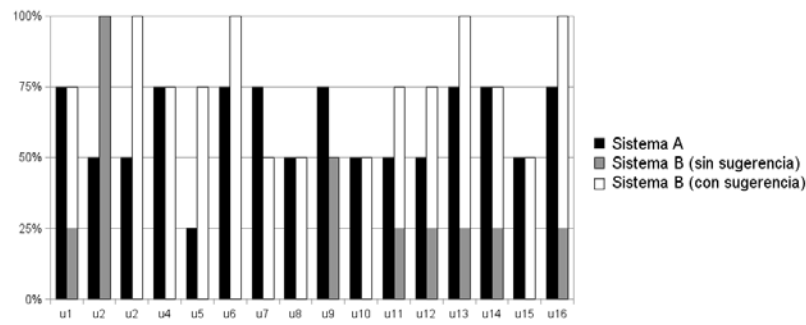


Figura 2. Precisión en sentido relajado para cada usuario.

para ser resueltas, fueron correctas 51, es decir, el 58,6%. Esto significa que la sugerencia de términos produce más éxitos. Este resultado no es igual para todos los usuarios: en la Fig. 2 pueden verse los resultados de manera individualizada.

En cuanto al tiempo medio empleado para resolver las preguntas correctas (en sentido relajado), fue de 65 segundos para el *sistema A* y 72 para el *sistema B*. Este resultado no debe engañarnos, pues se trata de valores medios para todo el *sistema B*. Aquellas preguntas que utilizaron la sugerencia de términos se resolvieron en realidad más rápidamente, con una media de 64 segundos. Habíamos esperado que la sugerencia de términos agilizase el proceso de búsqueda, no obstante, hay que tener en cuenta que el usuario requiere cierto tiempo para que pueda ver los términos sugeridos y seleccionar aquellos que considera más pertinentes para añadir a la búsqueda y resolver la pregunta.

4.2. Dificultad de las preguntas. Refinamiento de búsquedas

No todas las preguntas tuvieron la misma dificultad para ser resueltas. La Figura 3 muestra la precisión en sentido relajado de cada pregunta, independientemente del sistema utilizado. La pregunta 5, *¿Cuánto costó el Túnel del Canal?*, fue mal contestada en la mayoría de los casos porque en uno de los pasajes se

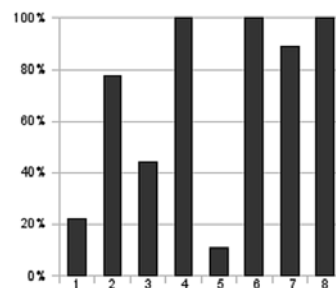


Figura 3. Precisión de las respuestas.

citaba una parte del coste de dicho canal, y no el coste completo. La pregunta 1, *¿Qué institución inició la campaña europea de la juventud contra el racismo?*, resultó difícil porque, a pesar de que la respuesta aparecía con poca dificultad, los usuarios desconocían la organización política de la UE en los años 1994 y 1995. La pregunta 3, *¿Quién era primer ministro de Noruega cuando se celebró el referéndum sobre su posible incorporación a la UE?*, resultó también complicada porque se trataba de una mujer, en vez de un hombre (recordemos que el sistema no aplicaba lematización en el proceso de indización).

Es importante analizar el refinamiento de las consultas. El número medio de consultas por pregunta fue de 3,6 para el *sistema A* y 3,4 para el *sistema B*, es decir, ligeramente inferior. Ahora bien, debemos tener en cuenta que los usuarios utilizaron la sugerencia de términos aproximadamente una vez por pregunta, y eso significa que de manera invisible para el usuario el motor de sugerencia tuvo que realizar una consulta. Hay que señalar también que en la mayoría de los casos la sugerencia de términos fue lanzada al comienzo de la búsqueda de las respuestas, un 67% de veces frente a un 33% que se lanzó en mitad de una búsqueda. Hemos comprobado, al igual que en otro estudio [4], que varios usuarios utilizaban la reformulación como mecanismo de verificación de respuestas, al incluir los términos de las mismas en la nueva reformulación, y eso puede dar lugar a errores en la interpretación.

4.3. Cuestionarios

Cada usuario respondió un cuestionario después de las sesiones en cada uno de los sistemas. El resultado se sintetiza en la Figura 4. En todos los aspectos, el *sistema B* fue considerado ligeramente mejor que el *sistema A*. En el cuestionario último que comparaba ambos sistemas y recogía comentarios de los usuarios, estos se decantaron claramente por el *sistema B* como mejor sistema. Un alto porcentaje de usuarios manifestaron que la sugerencia de términos les había presentado términos que de otro modo no habrían utilizado en sus consultas.

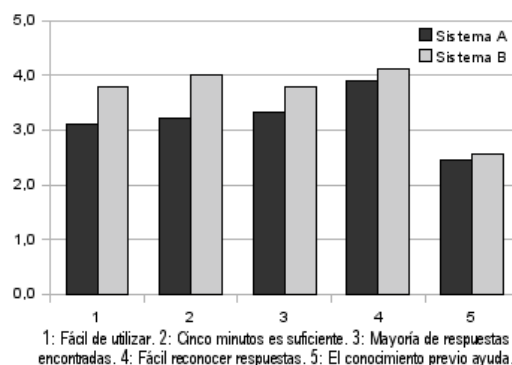


Figura 4. Resultados de los cuestionarios post-sistema.

5. Comentarios y conclusiones

Hemos explorado la actuación del usuario en un sistema interactivo de búsqueda de respuestas con sugerencia de términos. Para ello se ha utilizado un sistema convencional de recuperación de información basado en el modelo del espacio vectorial, al que se le han añadido ciertas características adicionales para la tarea interactiva: división de los documentos en pasajes de texto, sugerencia de términos para la expansión de la pregunta y una interfaz de usuario adecuada.

Uno de los primeros aspectos que deseamos resaltar es el hecho de haber obtenido una precisión del 68 % en la resolución de todas las preguntas, un valor muy alto. Esto indica que cuando la interacción es posible, un sistema simple proporciona mejores resultados que un sistema completamente automático [3].

En relación con el método interactivo de ayuda de la sugerencia de términos (obtenidos de acuerdo a la construcción automática de un tesoro de asociación), los resultados muestran que se trata de una opción apreciada por los usuarios, al permitir ver y utilizar términos que a priori el usuario no hubiese utilizado, obteniéndose mejores resultados.

Uno de los aspectos que los usuarios indicaron como muy útiles para la localización de la respuesta fue el resaltado de los términos de búsqueda en los pasajes de texto, pues rápidamente les permitía situarse en las zonas más prometedoras para encontrar la solución a las preguntas.

En relación con el sistema de recuperación en sí, con o sin sugerencia de términos, la mayoría de usuarios manifestó que era sencillo de utilizar, al parecerse mucho la interfaz de consulta a la de los motores de búsqueda en Internet. Alguno incluso llegó a indicar que la posibilidad de utilizar truncamiento, como en nuestro caso, era una opción muy acertada, no olvidemos que los usuarios procedían del campo documental.

Un aspecto que debemos comentar es si el experimento está bien planteado. Por ejemplo, los resultados con usuarios reales son difíciles de repetir, sobre todo si el número de usuarios es pequeño, o si cambia cualquiera de las condiciones de formación de los mismos (alumnos de otras titulaciones, por ejemplo). En este sentido es muy difícil incorporar a la evaluación el conocimiento previo que los usuarios pudieran tener en lo referente a sistemas de búsqueda de información, planteamiento de estrategias de búsqueda, conocimiento del propio sistema de recuperación, el bagaje cultural sobre las propias preguntas, o cómo son los procesos de inferencia que las personas utilizamos para resolver preguntas. El orden pseudo-aleatorio que se ha seguido ha pretendido reducir efectos no deseados en el orden de sistemas y preguntas, pero es seguro que no los ha eliminado por completo. También es difícil justificar por qué aparentemente algunas preguntas fáciles de responder por la mayoría de usuarios, a unos pocos les resultan especialmente complicadas.

Sería deseable repetir los experimentos con muchos más grupos de usuarios y de preguntas, pero los experimentos interactivos como el que hemos realizado suponen un duro trabajo de planificación, de selección de usuarios y en el desarrollo de los mismos, y muchas veces no es posible realizarlos. No obstante, es nuestra intención seguir con esta línea de investigación, y ya estamos analizando

técnicas de extracción de información y resumen automático que nos permita sugerir algo parecido a *conceptos* en vez de términos.

Agradecimientos

Este trabajo ha sido financiado por los proyectos KA3Ñ-2009 de la Universidad de Salamanca y FS/9-2008 de la Fundación Memoria Samuel Solórzano Barruso.

Referencias

1. Braschler, M., Peters, C.: CLEF 2003 methodology and metrics. In: CLEF 2003, LNCS. vol. 3237, pp. 7–20 (2004)
2. Figuerola, C.G., Zazo, A., Alonso Berrocal, J.L., Rodríguez, E.: Interactive and bilingual question answering using term suggestion and passage retrieval. In: CLEF 2004, LNCS. vol. 3491, pp. 363–370 (2005)
3. Gonzalo, J., Oard, D.W.: Overview of the CLEF 2004 interactive track. In: CLEF 2004, LNCS. vol. 3491, pp. 310–322 (2005)
4. He, D., Wang, J., Luo, J., Oard, D.W.: iCLEF 2004 at Maryland: Summarization design for interactive cross-language question answering. In: Working Notes for the CLEF 2004 Workshop, 15-17 September, Bath, UK (2004)
5. Hersh, W., Over, P.: The TREC-9 interactive track report. In: The Ninth Text REtrieval Conference (TREC 9). pp. 41–49 (2000)
6. Hersh, W., Over, P.: TREC-2001 interactive track report. In: The Tenth Text REtrieval Conference (TREC 2001). pp. 41–49 (2001)
7. López-Ostenero, F., Peinado, V., Gonzalo, J., Verdejo, F.: Interactive question answering: Is cross-language harder than monolingual searching? *Information Processing & Management* 44(1), 66 – 81 (2008)
8. Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F., de Rijke, M.: The multiple language Question Answering track at CLEF 2003. In: CLEF 2003, LNCS. vol. 3237, pp. 471–486 (2004)
9. Peñas, A., Gonzalo, J., Verdejo, F.: Acceso a la información mediante exploración de sintagmas. In: Canós, J., García, P. (eds.) III Jornada de Bibliotecas Digitales JBIDI'02. 18 y 19 de Noviembre. El Escorial (Madrid). pp. 57–66 (2002)
10. Salton, G., McGill, M.J.: *Introduction to Modern Information Retrieval*. McGraw-Hill, New-York (1983)
11. Spärck Jones, K.: A statistical interpretation of term specificity and its applications in retrieval. *Journal of Documentation* 28(1), 11–21 (1972)
12. Vicedo González, J.L.: *Recuperación de Información de Alta Precisión: Los Sistemas de Búsqueda de Respuestas*. Sociedad Española para el Procesamiento del Lenguaje Natural, Alicante (2003)
13. Woods, W.A., Green, S., Martin, P., Houston, A.: Halfway to question answering. In: The Ninth Text REtrieval Conference (TREC 9). pp. 489–500 (2000)
14. Zazo, A., Figuerola, C.G., Alonso Berrocal, J.L., Fernández Marcial, V.: Use of free on-line machine translation for interactive cross-language question answering. In: CLEF2005, LNCS. vol. 4022, pp. 263–272 (2006)
15. Zazo, A., Figuerola, C.G., Berrocal, J.L.A., Rodríguez, E., Gómez, R.: Experiments in term expansion using thesauri in Spanish. In: CLEF 2002, LNCS. vol. 2785, pp. 301–310 (2003)