

Evaluación de sistemas de recuperación de información web sobre dominios restringidos

Evaluation of Web Information Retrieval Systems on Restricted Domains

Javi Fernández
Universidad de Alicante
avifm@dsi.ua.es

José M. Gómez
Universidad de Alicante
mgomez@dsi.ua.es

Patricio Martínez-Barco
Universidad de Alicante
patricio@dsi.ua.es

Resumen Tradicionalmente en tareas de recuperación de información se han utilizado los métodos de evaluación definidos en foros internacionales como el R-C o el CL. Dichos métodos se basan principalmente en la cobertura y precisión a partir de los juicios de relevancia de las primeras n respuestas de cada sistema para un conjunto de consultas iniciales. En la práctica cuando un buscador se aplica a la web los usuarios no quieren una alta cobertura sino una alta precisión para obtener páginas muy relevantes en las primeras posiciones. Por ello que hemos adaptado los métodos de evaluación de estos foros valorando más la precisión que la cobertura. Finalmente se ha puesto en práctica este nuevo método de evaluación comparando dos buscadores web sobre un dominio restringido: oportunidades de negocio y exportación de productos.

Palabras clave evaluación recuperación de información web

Abstract Traditionally information retrieval systems are assessed using the same evaluation methods as proposed in the international forums like R-C or CL. Such methods mainly focus on the recall and precision obtained from the relevance judgements of the first n answers each system provides for a set of initial queries. In practice when applying a search engine to the web users do not want a high recall but a high precision in order to obtain really relevant pages at the first places. Therefore in this paper we have adapted the evaluation methods of these forums so that precision is given more importance than recall. To test this new evaluation method two web search engines are compared within the following restricted domains: business opportunities and product exports.

Keywords evaluation information retrieval web

1. Introducción

La recuperación de información (RI) es la ciencia que se encarga de la representación, almacenamiento, organización y obtención de información existente en diversas fuentes de datos no estructuradas. Los modernos motores de búsqueda en la web han aplicado muchas de las técnicas de la RI tradicional pero se han encontrado con nuevos problemas a resolver: la enorme cantidad de documentos existentes en la web (Vulli y Signorini 2002) su naturaleza dinámica (Toulas, Cho y Iston 2002) y el hecho de que una gran parte de la información no es accesible públicamente (Roder et al. 2002). Estos problemas hacen imposible conocer en un cierto instante de tiempo la totalidad de documentos existentes y el estado de cada uno de ellos. Pero al mismo tiempo la web posee una gran ventaja: la gran mayoría de documentos que contiene están relacionados entre sí mediante interconexiones formando un grafo (Roder et al. 2002) por lo que podemos acceder a gran cantidad de documentos mediante un conjunto reducido de semillas.

En la actualidad no existen métodos aceptados para la evaluación de este tipo de sistemas de RI (Sahami et al. 2002) por lo que las evaluaciones realizadas miden otro tipo de características

como el número de documentos indexados (Lawling et al. 2001) la velocidad de recuperación e indexación y la capacidad de eliminar duplicados y spam entre otros factores de naturaleza técnica más científica. De esta forma no es posible evaluar si un sistema es capaz de responder correctamente con resultados relevantes para una consulta concreta.

Lo más habitual para la evaluación de sistemas de RI es utilizar métodos inspirados en el R-C¹ o el CL². Lawling et al. (2001) realizan una evaluación de varios motores de búsqueda en la web utilizando *juicios de relevancia*. Para la evaluación de nuestro sistema utilizaremos una metodología similar.

Cabe destacar que el sistema que hemos desarrollado ha sido diseñado con el objetivo de ofrecer a los usuarios un conjunto de resultados lo más fiables posibles prefiriendo una alta *precisión* y dando menos importancia a la *cobertura* del sistema. Este interés es compartido por la mayoría de usuarios web a los que les interesan encontrar un documento que cumpla con sus requerimientos de información en las primeras posiciones. Según Silverstein et al. (1992) es extraño que un usuario

¹<http://trec.nist.gov>

²<http://www.clef-campaign.org>

mire más allá de los 10 primeros resultados. Si a esto sumamos el alto coste de medir la cobertura de un sistema de RI en la web debido al tamaño y dinamismo de Internet podremos afirmar que la precisión en los primeros n documentos es una medida suficientemente buena para evaluar este tipo de sistemas.

En la sección 2 describiremos los objetivos y las características del sistema de RI web desarrollado. En la sección 3 expondremos detalladamente el corpus utilizado, la metodología de evaluación aplicada y los resultados obtenidos tras la experimentación. Finalmente en la sección 4 haremos una breve reflexión sobre los resultados y plantearemos futuras mejoras tanto para el sistema de evaluación como para el sistema de RI.

2. Descripción del sistema

El sistema de RI seleccionado para realizar la evaluación es un sistema desarrollado por el grupo de procesamiento del lenguaje natural y sistemas de información³ LSI. Este sistema ha sido creado con el objetivo de ayudar a empresas a realizar las tareas de búsqueda en la web de una manera más rápida y fiable que con los motores de búsqueda actuales. Esta mejora se basa en dar prioridad a la *precisión* ofreciendo una cantidad menor de resultados rápidos pero con una mayor relevancia respecto a las consultas. Fiabilidad. Trabaja con un conjunto restringido de dominios previamente elegidos por el usuario.

Nuestro sistema está basado en una versión modificada de Lucene⁴. En esta versión tanto los términos de la consulta como los términos de los documentos indexados son analizados utilizando un *stemmer* y pesados adecuadamente. De esta forma conseguimos devolver un mayor número de documentos pero siempre dando más peso a aquellos que contienen los términos idénticos a los de la consulta original.

Por otra parte el pesado de documentos en el ranking de resultados ha sido diseñado con el fin de potenciar ciertas características que para los usuarios pueden ser importantes como por ejemplo mayor prioridad a los documentos más recientes y eliminación de aquellos más antiguos caducados utilizando un detector de fechas que devuelve la fecha más reciente dentro de cada documento y eliminación de duplicados o resultados muy similares que aparecen juntos en el ranking. Como valor añadido se han implementado ciertas características externas a la propia RI que pueden ayudar a disminuir el tiempo empleado en las búsquedas en la web: agrupación de documentos mediante clasificación automática y búsqueda en otros idiomas mediante un traductor externo.

El sistema además integra un *sistema de creación de juicios de relevancia* junto con los resultados de búsqueda. De esta forma los usuarios

son capaces de crear de forma sencilla y dinámica los juicios de relevancia para la evaluación al mismo tiempo que van realizando su trabajo. Así el tiempo consumido para esta tarea se minimiza. La utilización de este sistema de evaluación es muy simple añadiendo a cada resultado tres opciones para la evaluación: *marcar como relevante*, *marcar como no relevante* y *marcar como caducado*. Los dos primeros afectan al propio sistema de RI mientras que el último evalúa a la tarea realizada por el detector de fechas. Los resultados etiquetados se almacenan finalmente en una base de datos indicando de qué consulta provienen.

Con el fin de ayudar un poco más a los usuarios aprovechamos esta información para que aquellos documentos marcados como relevantes se coloquen en los primeros puestos del ranking en futuras búsquedas. Los no relevantes y los caducados se colocaran al final de la lista.

En la figura 1 se puede observar la integración de algunos de los elementos descritos anteriormente a la interfaz: las diferentes agrupaciones de resultados según motor de búsqueda, tipo de clasificación y categoría a la derecha los resultados donde el icono verde indica un resultado relevante y el rojo uno no relevante.

3. Evaluación

El sistema ha sido desarrollado con el fin de acelerar el trabajo de búsqueda en la web por lo que es preferible que el ranking contenga un número pequeño de resultados pero que todos ellos sean relevantes. Debido a eso y a que se trata de un sistema de RI en la web y por tanto no conocemos a priori el total de documentos existentes hemos considerado centrar la evaluación del sistema en medidas de *precisión*.

3.1. Corpus

El corpus utilizado consiste en un conjunto de documentos obtenidos mediante una herramienta de *crawling* desarrollada por el grupo pertenecientes a un conjunto restringido de dominios web dentro del sector de la exportación de productos. Aun que en la actualidad el corpus está en continuo crecimiento y actualización para el presente proceso de evaluación se ha utilizado una versión del mismo en un instante de tiempo concreto con el fin de trabajar con un corpus estático. Esta versión está formada por más de 10 millones de documentos de tipo *html* y *pdf* ocupando alrededor de 100 GB. El tamaño medio de los documentos es de unos 10 KB.

Como evaluación adicional hemos decidido realizar una comparación con el buscador en dominios restringidos que ofrece la empresa Google⁵. Mediante su herramienta *Google Custom Search*⁶ se ha creado un buscador personalizado con los mismos dominios indexados por nuestro

³<http://gplsi.dlsi.ua.es>

⁴<http://lucene.apache.org>

⁵<http://www.google.es>

⁶<http://www.google.com/cse>



Figura 1 estructura de la página de resultados de la herramienta

sistema. Se harán las mismas evaluaciones en ambos buscadores de manera paralela.

Realizaremos dichas evaluaciones con las 10 consultas más repetidas durante el tiempo de vida de este sistema realizadas por usuarios reales miembros de empresas del sector que han probado nuestro sistema. La longitud media de esas consultas es de alrededor de 2 términos.

3.2. Metodología

El sistema se ha evaluado utilizando juicios de relevancia. Estos juicios se han obtenido mediante la herramienta integrada ya mencionada en la sección 2. Para poder conseguir un conjunto de juicios suficiente para evaluar el sistema los miembros del grupo han colaborado durante una semana en la creación de los juicios de relevancia evaluando cada uno de ellos 2 de las 10 consultas mencionadas. Cada consulta ha sido evaluada por uno de los miembros utilizando su criterio como no expertos en el dominio.

3.3. Resultados

Nuestro objetivo principal desde las primeras fases de diseño del sistema es la *precisión* por lo que a la hora de extraer los resultados utilizaremos diferentes medidas que la cuantifiquen. Las medidas elegidas son *precisión al primer documento* (1 a los 5 primeros) y *precisión media* (MA) medidas usualmente utilizadas en los foros de evaluación Voorhees y Garman [2].

En el cuadro 1 exponemos los resultados generales de ambos sistemas. Podemos observar que en un 30% de los casos se obtiene un documento relevante como primer resultado utilizando nuestro sistema mientras que utilizando el buscador de Google obtenemos un 30%. Esta diferencia decrece ligeramente a medida que evaluamos la precisión a los 5 y a los 10. Si observamos el MA nuestro sistema me iguala al de Google en un 20%.

En el cuadro 2 exponemos los resultados de ambos buscadores para cada usuario que ha participado en la evaluación. Podemos ver que pese a las diferencias subjetivas de cada evaluador nuestro sistema ha mejorado al buscador perso-

Sistema	1	5	10	MA
Google	.3	.1	.1	.1
LSI	.1	.1	.3	.1

Cuadro 1 Resultados de precisión por sistema

nalizado de Google en un 30% y por tanto también en MA. Además utilizando nuestro buscador con el evaluador más pesimista la mitad de las veces encontramos un documento relevante en la primera posición del ranking. Cabe destacar que la diferencia de puntuación entre ambos buscadores permanece aproximadamente igual pese a las diferencias de criterios entre los evaluadores. Esto parece indicar que los criterios utilizados son independientes del sistema a evaluar.

Eval	Sistema	1	5	10	MA
1	Google	.1	.22	.2	.21
	LSI	.1	.32	.2	.332
2	Google	.1	.1	.1	.2
	LSI	.1	.1	.1	.1
3	Google	.1	.1	.1	.3
	LSI	.1	.1	.1	.1
	Google	.1	.1	.3	.1
	LSI	.1	.1	.2	.1
	Google	.1	.1	.1	.1
	LSI	.1	.1	.1	.1

Cuadro 2 Resultados de precisión de cada sistema por evaluador

En el cuadro 3 mostramos los resultados de ambos sistemas según el número de palabras de las consultas. En este cuadro se puede observar que es más sencillo encontrar documentos relevantes cuando las consultas tienen una longitud menor. Esto puede ser debido a que los evaluadores no conocen las verdaderas necesidades de información de los usuarios que hicieron las consultas al sistema y a que las consultas cortas son más generales. Cabe mencionar la existencia de

un repunte de precision cuando las consultas tienen cuatro palabras o mas pero esto puede ue ocurra por ue existen pocas consultas de este tipo y por lo tanto la muestra no es representativa.

al	Sistema	1	1	1	MA
1	oogle LSI 2 .	. . 22
2	oogle LSI	.222 12 . 32	. . 3
3	oogle LSI	.2 2 .	.3 1 .3 1	.3 .3	.332 .3 1
	oogle LSI	.333 .	. 33 . 33	. 1 .	. . 3

Cuadro 3 Resultados de precision de cada sistema por numero de palabras por consulta al

4. Conclusiones

uestra aproximacion midiendo unicamente la precision ha facilitado enormemente la tarea de evaluacion. sto es debido a ue para obtener una buena medida de cobertura habra sido necesario evaluar cientos de resultados por consulta en lugar de unicamente los 1 primeros del ranking midiendo solo la precision. Ademas esta forma de evaluacion se acerca mas a lo ue los internautas demandan ue consiste en obtener documentos relevantes en las primeras posiciones.

n este art culo hemos pretendido mostrar como se pueden evaluar los motores de bus ueda en la eb usando la precision. ara ello hemos comparado un sistema de RI web sobre dominios restringidos desarrollado en el LSI con el buscador personaliado de oogle. Los resultados obtenidos nos alientan a seguir esta lnea de investigacion y me orar el sistema a partir de los comentarios de los usuarios y evaluadores.

ese a las facilidades ue nuestra interfa ha ofrecido a los evaluadores medir este tipo de sistemas sigue siendo un problema arduo y tedioso. s por ello ue pretendemos me orar la interfa con el fin de aportar mayor informacion al usuario y ue este sea capa de tomar una decision de la relevancia o no de cada resultado sin necesidad de visitar el documento. uestra primera aproximacion a este problema sera me orar la seleccion de los snippets centrandolos en las partes verdaderamente importantes del documento. ara ello utilizaremos modelos basados en densidad de terminos relevantes y modelos de n-gramas.

ebido al dinamismo de la eb nos hemos enfrentado a una serie de dificultades añadidas a la hora de evaluar. stos problemas se han planteado debido a ue algunas veces los contenidos de las paginas cambian y los documentos ue eran relevantes de aron de serlo o viceversa. sto implica repetir algunos experimentos. ara futuras

evaluaciones pretendemos almacenar el contenido de los documentos en el momento de la evaluacion con el fin de evitar estos problemas.

or otra parte pretendemos me orar la precision del sistema del LSI incorporando los modelos de densidad y n-gramas mencionados arriba as como eliminar el contenido de los documentos web ue no aportan informacion como pudieran ser menus cabeceras pies de pagina etc. tro aspecto a me orar en nuestro buscador es el metodo de pesado de terminos basado en *tf/idf*. Queremos probar otros modelos de pesado para encontrar el ue me or precision nos aporte.

5. Agradecimientos

ste art culo ha sido cofinanciado por el MICI proyecto I 2 -133 1-C - 1 y la Conselleria d educacion de la generalitat Valenciana proyectos R M 2 11 y AC M 2 1 2 .

Bibliografía

roder Andrei Ravi umar ar in Maghoul rabha ar Raghavan Sridhar Ra agopalan Raymie Stata Andrew om ins y Janet iener. 2 . raph structure in the web. n *Proceedings of the 9th international WWW conference on Computer networks* paginas 3 32 . orth-olland ublishing Co.

ulli Antonio y Alessio Signorini. 2 . he indexable web is more than 11. billion pages. n *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web* paginas 2 3. ACM.

aw ing avid ic Craswell eter ailey y athleen ri hs. 2 1. Measuring search engine uality. *Information Retrieval* 1 33 .

toulas Alexandros Junghoo Cho y Christopher lston. 2 . hat s new on the web? the evolution of the web from a search engine perspective. n *WWW '04: Proceedings of the 13th international conference on World Wide Web* paginas 1 12. ACM.

Sahami Mehran Vibhu Mittal Shumeet alu a y enry Rowley. 2 . he happy searcher Challenges in web information retrieval. n *PRICAI 2004: Trends in Artificial Intelligence: 8th Pacific Rim International Conference on Artificial Intelligence* volumen 31 paginas 3 12. Springer Verlag.

Silverstein Craig annes Marais Moni a enger y Michael Moric . 1 . Analysis of a very large web search engine uery log. *SIGIR Forum* 33 1 12.

Voorhees llen M. y onna . arman. 2 . *TREC: Experiment and Evaluation in Information Retrieval*. igital Libraries and lectronic ublishing. MI ress.