Mejoras en la recuperación web combinando campos

Carlos G. Figuerola, José Luis Alonso Berrocal, y Angel Zazo Rodríguez

Grupo de Investigación REINA, Universidad de Salamanca
{figue, berrocal, zazo}@usal.es

Resumen. Este artículo describe algunas de las actividades del grupo de investigación REINA en torno a la recuperación de información web. Estas actividades se han centrado en probar la capacidad de recuperación que puede esperarse de diversos elementos informativos presentes en las páginas web, además del texto que el usuario visualiza normalmente en su navegador. Nuestro objetivo ha sido probar estrategias pre-recuperación de mezclar o combinar esos campos o elementos de información. Combinar términos de diversa procedencia en un único índice puede conseguirse, en sistemas basados en el modelo del espacio vectorial, operando sobre la fecuencia del término en el documento, si se aplica un esquema de pesado basado en $tf \times idf$. El campo `BODY` es, obviamente, el más potente desde el punto de vista de la recuperación; pero los `ANCHOR`s de los *backlinks* que reciben las páginas indizadas añaden una mejora considerable a los resultados de la recuperación. El contenido de las etiquetas `META`, sin embargo, contribuyen poco a la mejora en la recuperación.

Palabras clave: Recuperación de información, recuperación web, modelo vectorial

1 Introducción

Este artículo describe los trabajos efectuados por el grupo de Investigación REINA en torno a la posible mejora de los resultados en la recuperación de páginas web en base a la consideración de diversos campos o elementos de dichas páginas. La colección de documentos utilizada como base ha sido la conocida como EuroGov, utilizada ampliamente en trabajos experimentales de este tipo [10]. En realidad, nuestros trabajos han utilizado la parte de documentos que están en castellano; tanto los que pertencen al dominio .es, como los que, perteneciendo a otros dominios diferentes, están tambien en esta lengua.

De otro lado, muchos documentos pertenecientes al dominio .es están en otras lenguas. Además de las otras lenguas que también se hablan en España (catalán, euskera, gallego), muchas de las páginas tienen versiones o traduciones internacionales; especialmente en inglés, pero también en otras como francés o alemán. Adicionalmente, y por parecidas razones, páginas pertencientes a otros dominios de EuroGov tambiés estaban en castellano, por lo que se decidió explorar la colección entera recolectando todas las páginas cuyo idioma fuera el castellano.

Lamentablemente, las cabeceras de las páginas web ofrecen información muy poco fiable; tanto sobre el idioma de las mismas como sobre otras cosas. En muchas ocasiones esas cabeceras están vacías; mientras que en otros tienen contenidos que no se corresponden con la realidad. Así pues, tuvimos que recurrir a un detector de lenguaje para seleccionar las páginas en castellano existentes en la colección EuroGov. El detector utilizado es *TextCat* [7], un programa basado en la construcción de patrones de uso de *n-gramas* para cada lengua contemplada; y la posterior categorización del texto cuya lengua se desea averiguar [2].

Por lo que se refiere a las consultas, se utilizaron las elaboradas en castellano durante las campañas CLEF 2005 y 2006 [11], junto con los correspondientes juicios de relevancia.

Este artículo está organizado de la siguiente manera: en la próxima sección se describe el enfoque adoptado para abordar el problema; en la sección siguiente se exponen los problemas

y las opciones adoptadas en lo que se refiere a análisis léxico y extracción de términos de las páginas. A continuación, se discute acerca de los campos o elementos de información útiles en la indización y recuperación. Más adelante se muestran los experimentos llevados a cabo y sus resultados. Finalmente, se extraen unas conclusiones.

2 Nuestro enfoque

La estrategia básica seguida es: primero, encontrar las páginas más relevantes para cada consulta y, de acuerdo con el tipo de consulta, reordenar la lista de los documentos más relevantes recuperados. Así, nuestro trabajo tiene dos partes fundamentales: encontrar páginas relevantes y después ordenar de una forma útil esas páginas. La tarea de recuperación de las páginas relevantes para una consulta determinada puede abordarse mediante un sistema convencional de indización y recuperación, como los basados en el modelo del espacio vectorial [9].

Sin embargo, en una página web hay, además del texto que el usuario ve en la ventana de su navegador, otros elementos informativos que pueden resultar interesantes desde el punto de vista de la recuperación. Incluso dentro del propio texto podemos encontrar ciertas estructuras que pueden ayudar, junto con esos otros elementos informativos, a mejorar los resultados de la recuperación.

Por ejemplo, algunos de esos otros lementos mencionados son el campo `TITLE`, algunas etiquetas `META`, los anclas de los *backlinks*, etc. Dentro del campo `body`, que es lo que se visualiza en la ventana del navegador, podemos diferenciar partes que usan diferentes tipografías, por ejemplo. Así, tenemos diferentes fuentes de información que podemos fusionar. En este sentido, se han propuesto dos estrategias básicas de fusión: fusión o combinación pre-recuperación y post-recuperación [1], [6]. Ambas estrategias han sido probadas; la estrategia post-recuperación consistió en construir un índice para cada elemento de información a combinar. Cuando hay que resolver una consulta, ésta es ejecutada contra cada uno de los índices construídos; los resultados obtenidos con cada índice son fusionados después en una única lista de documentos recuperados.

También hemos probado la estrategia pre-recuperación, que se ha concretado el la elaboración de un único índice con los términos de todos los elementos o campos contemplados; pero pesado de manera diferente. Una vez construído este único índice, las consultas son ejecutadas normalmente contra él. Naturalmente, en la elaboración de este único índice podemos buscar dar más o menos valor a un término en función de que aparezca en un campo o en otro. Esto nos permite utilizar una combinación que, bien afinada, debería proporcionar buenos resultados en la recuperación.

La aplicación de diferente peso a los diferentes componentes de la mezcla, en nuestro caso, es fácil. Para la indización y recuperación utilizamos nuestro software *Karpanta* [4], basado en el conocido modelo vectorial, y, en consecuencia, no tenemos más que operar sobre la frecuencia de cada término en cada uno de los componentes de la combinación.

En este caso hemols aplicado un esquema de pesado `ATU` (slope=0.2) [12], pero la idea es similar para cualquier esquema de pesado basado en $tf \times idf$. Podemos aplicar un multiplicador a tf en función del campo en que aparezca el término, de manera que haga aumentar o disminuir el peso de dicho término, sin dejar de considerar la frecuencia del término y el IDF.

2.1 Análisis léxico y extracción de términos

La primera opreación, previa a cualquier otra, es la conversión de las páginas web a texto plano. Esto también era necesario para detectar la lengua de cada página, dado que se deciió trabajar sólo con las que estaban en castellano.La obtención del texto plano no es

trivial, y no está exenta de problemas. Como se ha comentado antes, no se puede confiar que en que se sigan los estándares en todos los casos; ni siquiera es posible asegurar que el contenido sea HTML, aunque el documento comience con las etiquetas apropiadas. En algunas ocasiones podemos nos podemos encontrar directamento con código binario, PDFs y similares inmediatamente después de un `<HTML>`. Con las etiquetas `META` sucede lo mismo; incluso cuando están presentes, no siempre ofrecen información correcta.

De hecho, muchas páginas ni siquiera contienen texto; así que lo primero es determinar el tipo de contenido. El viejo y bien conocido comando *file*, bien afinado, puede ayudar a determinar el contenido real de cada página. Adicionalmente, nos informará sobre otra cuestión importante: el sistema de codificación de caracteres utilizado. Para páginas en castellano, lo habitual es *ISO 8859* o *UTF-8*; es necesario conocer esta información para tratar adecuadamente los caracteres especiales. La conversión a texto plano, cuando el documento parece ser realmente HTML, la efectuamos con *w3m*. Hay otros conversores, pero después de varias pruebas los mejores resultados parecen ser los de *w3m*.

Una vez obtenido el texto plano, podemos determinar el idioma mediante *TexCat*. Seleccionados los documentos en castellano, es preciso extraer términos y normalizarlos en alguna manera. Básicamente, los caracteres son convertidos a minúsculas, los acentos son eliminados, al igual que números, signos ortográficos y similares. Se eliminan también palabras vacías, de acuerdo con una lista estándar para el castellano; y los demás términos se pasan a través de un s-stemmer mejorado [5].

2.2 Campos utilizados

Son varios los elementos o fuentes de información que podemos considerar en una página web. La base es el campo `BODY`, obviamente, pero además podemos usar el campo `TITLE`, que parece claramente descriptivo, al igual que varias etiquetas `META` que habitualmente son utilizadas para estos propósitos, en especial `META content="Description"` and `META content="Keywords"`. Sin embargo, como ya hemops indicado en otro lugar ([3]), estos campos no están presentes de una manera uniforme en todas las páginas. Muchas no los contienen y otros tienen valores `META` generados de manera automática por los programas que han producido esas páginas web.

En otros casos, aunque los valores sean asignados de manera manual por los autores de las páginas, resultan poco operativos. Adicionalmente, podemos pensar que términos que aparecen con tipografía resaltada pueden ser más representativos. Las etiquetas o campos `H1`, `H2`, etc. son un ejemplo. Desgraciadamente, el uso de estas etiquetas tampoco es uniforme, y ha evolucionado hacia la definición de fuentes y tamaños específicos de letra, más complejos de procesar.

Otro elemento importante es el de los anclas de los *backlinks*, o enlaces desde otros lugares a la página en cuestión. De manera más o menos breve, estos anclas describen la página con la que conectan; esta descripción es importante, porque está hecha por alguien diferente del autor de la página que queremos indizar. Podemos pensar que esta descripción añade términos distintos de los usados en la propia página. No obstante, muchos de estos anclas pueden hacer referencia a partes muy concretas de la página apuntada; igualmente, hay páginas con muchos *backlinks* y anclas y otras con pocos o ninguno. Y, en cualquier caso, tenemos el problema de obtener esos anclas. En nuestro caso, hemos procesado la totalidad de la colección EuroGov para obtener todos los enlaces y sus correspondientes anclas que apuntan hacia páginas en castellano. Es evidente que fuera del la colección EuroGov hay más enlaces que apuntan hacia dichas páginas, pero no tenemos forma de obtenerlos.

Finalmente, hemos construido índices con los siguientes elementos: `BODY`, `TITLE`, `META content=Title`, `META content=description`, `META content=keywords`, `H1`, `H2`, `ANCHORS`.

3 Experimentos

Apoyándonos en las estimaciones de relevancia que forman parte de la colección de documentos, hemos efectuados diversas pruebas con varias combinaciones de elementos o campos en distintas proporciones. Hemos elegido como línea base para comparaciones los resultados con un índice formado exclusivamente por los términos que aparecen en el campo `BODY`.

`BODY`(fd=1)
`ANCHOR`(fd=1)
`TITLE` (fd=1.5)
`META-DESC` (fd=1.5)
`META-TITLE` (fd=1)
`META-KEY` (fd=0.5)
`H1` (fd=0.8)
`H2` (fd=0.8)

Tabla 1. Combinación ganadora

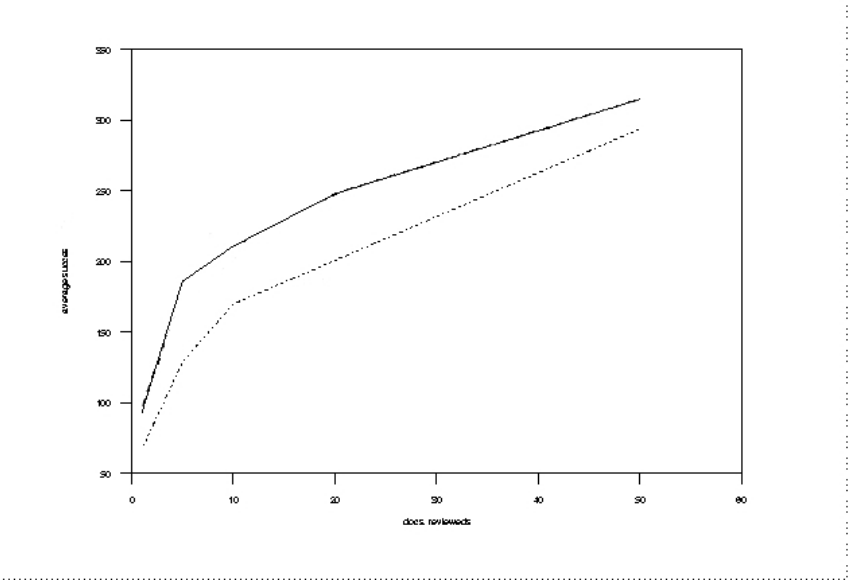

Fig. 1. Resultados con campo `BODY` y con Combinación Ganadora

Como era de esperar, el uso de cualquiera de esos elmentos en adición al campo `BODY` mejora los resultados de la recuperación. Pero algunos de esos elementos lo hacen de forma notable y otros no tanto.

También hemos efectuado pruebas con índices compuestos por uno solo de cada uno de los campos. Esto nos permite observar la utilidad, desde el punto de vista de la recuperación, de cada uno de esos campos. El gráfico muestra los resultados; cada curva representa los obtenidos tras ejecutar la batería de consultas contra un índice elaborado con los términos de cada campo o elemento de información ($fd = 1$). Cada curva se refiere a uno solo de esos campos, sin utilizar los términso del `BODY`. Los campos `H1` y `H2` no aparecen en el

gráfico porque producen resultados extremadamente pobres. Es evidente que muchas páginas carecen de uno o varios de estos campos, razón por la cual nunca podrán ser recuperados a partir de esos campos; pero es una de verificar por separado la capacidad de recuperación de esos campos.

Como era esperable, cada campo por separado produce peores resultados que `BODY`, lo cual es normal, pues este campo contiene la parte visualizable de la página web. Pero tras el campo `BODY` el campo que parece más interesante desde el punto de vista de la recuperación es el de los `ANCHOR`s de los *backlinks*, aunque en muchos casos tales anclas sean muy breves. Pero parece que, como se ha dicho, las descripciones que otros hacen de una página son muy eficaces para su recuperación.

Le sigue a corta distancia el campo `TITLE`, algo previsible. Sin embargo, los campos basados en etiquetas `META` ofrecen resultados bastante pobres, a bastante distancia de `ANCHOR` y `TITLE`; y eso a pesar de que son campos concebidos específicamente para la recuperación. Aunque hay poca diferencia entre los resultados de los tres `META` observados (`title`, `description` y `keywords`) es éste último, curiosamente, el que peores resultados produce de los tres.

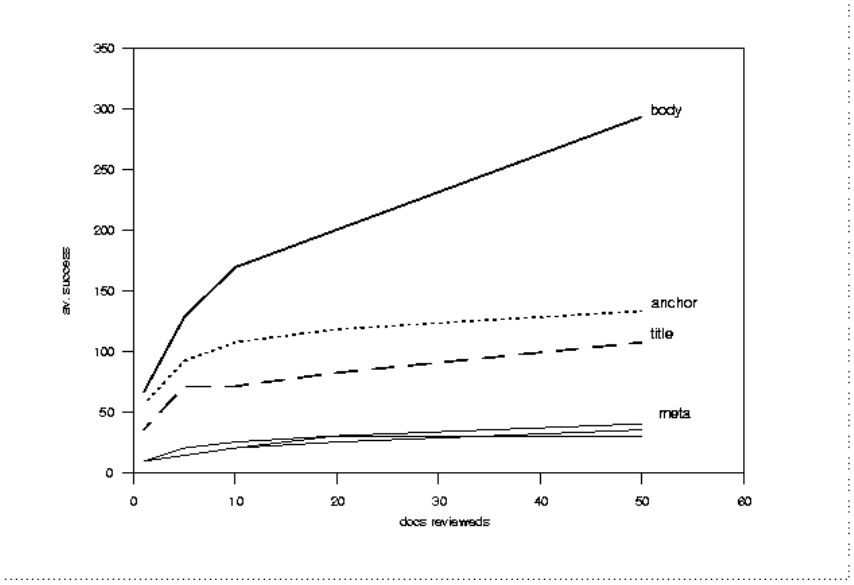

Fig. 2. Aportación de campos individuales

4 Conclusiones

Hemos descrito algunos de los experimentos realizados sobre la recuperación de páginas web a partir de determinados elementos. El uso de tales campos o elementos de información, además del texto dela zona `BODY` de las páginas web permite mejorar los resultados de la recuperación de información. Una forma de hacerlo es elaborar un único índice con los términos que aparecen en esos campos, junto con los que aparecen el el `BODY`; pero pesándolos de manera diferente, ajustable de forma empírica.

De todos esos campos, parece que el más eficaz es el de los `ANCHORS` de los *backlinks* que recibe cada página. El campo `TITLE` también contribuye de manera importante a la mejora

de la recuperación. El contenido de las etiquetas `META`, sin embargo, parece de utilidad reducida, desde el punto de vista de la recuperación.

Referencias

1. Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, Ophir Frieder, and Nazli Goharian. On fusion of effective retrieval strategies in the same information retrieval system. *Journal of the American Society for Information Science and Technology (JASIST)*, 55(10):859–868, 2004.
2. William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *Third Annual Symposium on Document Analysis and Information Retrieval. April 11-13, 1994, Las Vegas, Nevada*, pages 161–175, 1994.
3. Carlos G. Figuerola, José L. Alonso Berrocal, Ángel F. Zazo Rodríguez, and Emilio Rodríguez. REINA at the WebCLEF task: Combining evidences and link analysis. In Peters [8].
4. Carlos G. Figuerola, José Luis A. Alonso Berrocal, Ángel F. Zazo Rodríguez, and Emilio Rodríguez Vázquez de Aldana. Herramientas para la investigación en recuperación de información: Karpanta, un motor de búsqueda experimental. *Scire*, 10(2):51–62, 2004.
5. Carlos G. Figuerola, Ángel F. Zazo, Emilio Rodríguez Vázquez de Aldana, and José Luis Alonso Berrocal. La recuperación de información en español y la normalización de términos. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 8(22):135–145, 2004.
6. Edward A. Fox and Joseph A. Shaw. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*. NIST Special Publication 500-215, 1993.
7. Gertjan van Noord. Texcat language guesser.
8. Carol Peters, editor. *Results of the CLEF 2005 Cross-Language System Evaluation Campaign. Working notes for the CLEF 2005 Workshop, 21-23 September, Vienna, Austria*, 2005.
9. Gerard Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communication of the ACM*, 18:613–620, 1975.
10. B. Sigurbjornsson, J. Kamps, and M. De Rijke. EuroGOV: Engineering a multilingual Web corpus. *Lecture Notes in Computer Science*, 4022:825, 2006.
11. Börkur Sigurbjörnsson, Jaap Kamps, and Maarten de Rijke. Overview of webclef 2005. In Peters [8].
12. Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 18–22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)*, pages 21–29. ACM, 1996.