Informe Técnico - Technical Report
DPTOIA-IT-2002-007
Diciembre, 2002

Tesauros de asociación y similitud para la expansión automática de consultas. Algunos resultados experimentales.

Ángel F. Zazo Rodríguez
Carlos G.-Figuerola Paniagua
José Luis Alonso Berrocal
Emilio Rodríguez Vázquez de Aldana

Departamento de Informática y Automática
Universidad de Salamanca

Revisado por

Dr. Joaquín García Palacios
Departamento de Traducción e Interpretación
Universidad de Salamanca

Dr. Luis Antonio Miguel Quintales
Departamento de Informática y Automática
Universidad de Salamanca

Aprobado en el Consejo de Departamento de 2 de diciembre de 2002

Información de los autores:

Ángel F. Zazo Rodríguez
Dr. Carlos G.-Figuerola Paniagua
Dr. José Luis Alonso Berrocal
Emilio Rodríguez Vázquez de Aldana

Todos ellos son integrantes del
Grupo de Recuperación Automatizada de la Información (REINA)
Área de Lenguajes y Sistemas Informáticos.
Departamento de Informática y Automática.
Facultad de Traducción y Documentación. Universidad de Salamanca
C/ Francisco Vitoria, 6-16. 37008 - Salamanca
{afzazo,figue,berrocal,aldana}@usal.es
http://reina.usal.es

Resumen

En los sistemas de recuperación de información uno de los aspectos que más condiciona los resultados es la formulación de la consulta. Ello supone seleccionar los términos que semánticamente se ajusten más a la necesidad informativa del usuario. Lamentablemente, figuras lingüísticas como la polisemia y la sinonimia hacen que esta tarea no sea nada fácil. Las técnicas que intentan reducir en lo posible el problema parten generalmente de una primera consulta, y estudian cómo puede modificarse para obtener mejores resultados. Uno de los mecanismos utilizados se conoce como expansión automática de consultas. Esta expansión consiste, primero, en añadir nuevos términos a la consulta original, y segundo, calcular el peso de dichos términos en la nueva consulta. Para llevar a cabo la expansión de consultas se han planteado varios mecanismos. Uno de ellos es la utilización de un tesauro construido automáticamente a partir de la colección de documentos. En este trabajo hemos experimentado con dos tipos de tesauros: de asociación y de similitud. Estos tesauros recogen las relaciones entre los términos de la colección documental, y sirven para expandir los términos originales de la consulta con aquellos más relacionados. Uno de los aspectos más importantes de tal expansión consiste en determinar el peso de los términos expandidos. En este sentido, también hemos realizado varios experimentos que muestran la dependencia entre la elección del mecanismo de pesado y los resultados, así como la influencia de la normalización de la consulta original en los mismos.

Abstract

One important aspect in Information Retrieval is the formulation of the query. Lexical figures as the synonymy and polysemy cause that the same concept can be expressed with different words and the same word can appear in documents that deal with different topics. Many techniques have been used to try to reduce this problem, inter alia automatic query expansion. This technique involves two basic steps: expanding the original query with new terms, and reweighting the terms in the expanded query. Several approaches exist to carry out this task, one of the most important is the use of a thesaurus. This technical report shows the work carried out by our research group about query expansion using association and similarity thesaurus, specially in aspects related to the weight of the terms added to the query.

Índice

1. Introducción

Uno de los problemas más importantes en Recuperación de Información (RI) consiste en formular la consulta para que plasme adecuadamente la necesidad informativa del usuario. El mayor inconveniente consiste en determinar el conjunto de términos que expresen semánticamente esa necesidad. El problema se agrava debido al efecto de inconsistencia en la asignación subjetiva de términos a conceptos, o lo que es lo mismo, que dos personas utilizan palabras diferentes para definir los mismos conceptos [3]. Figuras como la sinonimia o la polisemia hacen que el mismo concepto pueda expresarse con palabras diferentes y una misma palabra puede aparecer en documentos que traten sobre temas distintos.

En esta situación, no es extraño que el usuario tenga que reformular su consulta para obtener mejores resultados. El usuario parte de una primera consulta que lanza al sistema de recuperación, y éste le devuelve un conjunto de documentos ordenados por algún criterio del sistema. El usuario, después de examinar esos documentos, tiene que reformular su consulta para obtener resultados más relevantes a su necesidad informativa. Se han planteado diversos mecanismos para construir la nueva consulta, pero en general, en todos ellos se realiza una ampliación de nuevos términos a la consulta inicial, y un recálculo del peso de los términos en la nueva consulta. Esto es lo que se conoce como expansión de consultas. Con esta expansión pueden mejorarse los resultados de búsqueda, si bien, en contraposición el coste computacional o los tiempos de respuesta pueden aumentar.

En los sistemas de recuperación de información que aplican pesos a los términos que representan a los documentos, el número de términos originales de la consulta influye en los resultados. En general, los resultados son mejores cuanto mayor es el número de términos de la consulta, pues se incluyen más términos índice que representan a los documentos relevantes [22]. Parece claro, en este sentido, que el interés para realizar la expansión se centra en consultas con muy pocos términos. Este tipo de consultas tienen un interés especial, pues son las que porcentualmente más se realizan en los motores de búsqueda de Internet, de uno a tres términos por consulta [6, 21].

Para expandir la consulta deben utilizarse palabras o frases con significado similar a aquellos de la consulta inicial. Existen diferentes estrategias para seleccionar nuevos términos para la consulta original. Uno de esos mecanismos es la utilización de un tesauro. Podría utilizarse un tesauro general para expandir la consulta, sin embargo, esto no suele dar buenos resultados [20], debido fundamentalmente a que las relaciones en un tesauro general no son válidas en el ámbito de la colección de documentos. Se obtienen mejores resultados si se utilizan tesauros o técnicas de expansión de consultas construidos a partir de la colección de documentos sobre la que se lanzan las búsquedas. Se parte de la hipótesis de que si dos documentos son similares, respecto de algún criterio, los términos que aparezcan frecuentemente en los mismos están a su vez relacionados. Cuando el tesauro se construye automáticamente, sin información adicional de relevancia del usuario, se distinguen varios enfoques [5]:

1. Tesauros construidos a partir de la medida simple de coocurrencias de términos [8]. La similitud entre términos se realiza basándose en la Hipótesis de Asociación: si un término es buen discriminante de documentos relevantes y no relevantes, sus términos asociados también lo serán [19].

2. Tesauros construidos a partir de *clustering* de documentos [2]. Primero se clasifican los documentos, y los términos poco frecuentes en una clase se utilizan para construir el tesauro de términos relacionados.

3. Tesauros de similitud construidos realizando la transposición de la matriz documentos-términos, es decir, representando los términos en función de los documentos [14, 11].

4. Tesauros construidos a partir de la asociación de términos y frases. Una frase es un conjunto de palabras que gramaticalmente satisfacen alguna regla del detector de frases (nombre-nombre-nombre, nombre-adjetivo, etc.). Se identifican frases en el texto y se asocian a términos [7].

5. Tesauros basados en información sintáctica. La relación entre términos se realiza en base a conocimiento sintáctico y análisis de coocurrencias. Se emplean gramáticas y diccionarios para obtener los términos relacionados con uno dado [4].

En este trabajo hemos experimentado los enfoques 1 y 3, pues son relativamente simples y efectivos. Para ello se ha utilizado una colección extensa de documentos en español. En la siguiente sección se indica cómo se construyen los tesauros con ambos enfoques. En la Sec. 3 se explican los experimentos realizados, se comparan los resultados y, finalmente, en la Sec. 4 se muestran las conclusiones.

2. Construcción automática de un tesauro

Un tesauro de términos es una matriz que mide relaciones entre los términos de la colección de documentos [16]. Esta matriz se utiliza para expandir los términos de la consulta con aquellos mejor relacionados. La matriz puede verse como una descripción semántica de términos, que refleja las influencias de unos términos sobre otros [1]. El tesauro T puede representarse utilizando una matriz cuadrada $n \times n$, como indica (1), siendo n el número de términos de la colección. La columna i representa la relación del término t_i con el resto de términos de la colección, $\vec{t_i} = (r_{i1}, r_{i1}, \ldots, r_{in})^T$, donde r_{ik} indica la relación del término t_k con el término t_i.

$$T = \begin{pmatrix} r_{11} & r_{21} & \ldots & r_{n1} \\ r_{12} & r_{22} & \ldots & r_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n} & r_{2n} & \ldots & r_{nn} \end{pmatrix} \tag{1}$$

Destacamos dos aspectos fundamentales a la hora de aplicar la información de la matriz en nuestros experimentos. Primero, la expansión se realiza con aquellos términos que están *mejor* relacionados con uno dado. Es decir, no se consideran valores umbrales a partir de los cuales un término se expanda con todos los relacionados. En vez de ello se seleccionan los términos que tengan los valores más altos de relación con cada término dado de la consulta original. Y segundo, se consideran los términos mejor relacionados con todos los términos de la consulta en general, y no con cada uno de ellos por separado.

Antes de continuar describiendo la construcción de los tesauros, debemos indicar que en la literatura sobre el tema los resultados obtenidos no han sido todo lo esperanzadores que cabría esperar. En [9] se ofrece quizás el estudio teórico más crítico de los modelos basados en coocurrencias. Un estudio previo de [17] muestra asimismo como la expansión de consultas con términos seleccionados al azar ofrece incluso mejores resultados que aquellos obtenidos del cómputo de coocurrencias. A pesar de ello, los resultados experimentales que nosotros hemos obtenido son altamente satisfactorios, quizás debido a las dos consideraciones sobre la expansión de la consulta que se han indicado previamente.

2.1. Matriz de asociación

La coocurrencia de términos se ha utilizado a menudo en RI para identificar términos que son similares a otros. La idea subyacente es que si dos términos aparecen frecuentemente en los mismos documentos estarán relacionados de alguna forma. La información obtenida del análisis

de coocurrencias puede utilizarse para establecer relaciones semánticas entre los términos. De hecho, se parte de la Hipótesis de Asociación [19, p. 104]. Considerando que los términos que aparecen en la consulta son buenos discriminantes de los documentos relevantes y no relevantes, también lo serán sus términos relacionados. Ello nos permite añadirlos a la consulta original.

Para medir el grado de asociación entre dos términos existen diferentes coeficientes. Todos ellos utilizan cálculos simples que miden el número de documentos en los que aparecen dos términos por separado, comparándolo de alguna manera, con el número de documentos en los que ambos coocurren:

$$\mathrm{ASO}(t_i, t_j) = f(c_i, c_j, c_{ij}) \tag{2}$$

siendo c_i y c_j el número de documentos en los que aparece el término t_i y t_j, respectivamente, y c_{ij} el número de documentos en los que coocurren t_i y t_j. Para nuestros experimentos hemos utilizado tres coeficientes habituales para la medida de similitud entre términos: Tanimoto, Coseno y Dice [14].

$$\begin{aligned} \mathrm{Tanimoto}(t_i, t_j) &= \frac{c_{ij}}{c_i + c_j - c_{ij}} \\ \mathrm{Coseno}(t_i, t_j) &= \frac{c_{ij}}{\sqrt{c_i \cdot c_j}} \\ \mathrm{Dice}(t_i, t_j) &= \frac{2 \cdot c_{ij}}{c_i + c_j} \end{aligned}$$

El valor de estos coeficientes está entre 0 y 1: si dos términos no coocurren en ningún documento el valor es cero; el valor más alto de asociación es 1 para aquellos términos que aparezcan únicamente en los mismos documentos. La matriz de asociación que se obtiene es una matriz simétrica, con la diagonal principal con valores unidad.

2.2. Matriz de similitud

En este apartado seguimos las explicaciones de [11, Sec. 3]. Un tesauro de similitud es una matriz que se construye utilizando relaciones entre términos. En su elaboración no se mide la coocurrencia de términos en los documentos, sino que se utiliza un mecanismo mediante el cual cada término de la colección se caracteriza por los documentos en que aparece. Es darle la vuelta al concepto clásico de los sistemas de recuperación de información. Para construir el tesauro de similitud, los términos de la colección se consideran documentos, y los documentos se utilizan como términos índice (*documentos índice*). Es decir, podemos pensar que los documentos pueden servir para representar los términos. Primero se representan los términos en función de los documentos, y segundo se computa su similitud para obtener el tesauro.

Tomemos como base el modelo vectorial [12], en el que generalmente se utiliza un esquema de pesado *tf-idf* [15]. En este modelo, dos documentos son similares si contienen los mismos términos, con más o menos, la misma frecuencia. Dando la vuelta a este modelo podemos pensar que dos términos son similares si aparecen en los mismos documentos.

Utilizando este esquema, al representar términos en función de los m documentos de la colección, cada término t_i vendrá representado por un vector de m componentes en el espacio vectorial de documentos, $\vec{t_i} = (p_{i1}, p_{i1}, \ldots, p_{im})$, siendo p_{ij} un valor que expresa el peso del documento índice d_j en la representación del término t_i. Para computar el valor de p_{ij} se utiliza el esquema de *tf-idf*, pero invirtiendo el papel de términos y documentos. El cálculo propuesto en [11] es el indicado en la Ec. (3). Es el que hemos utilizado en nuestros experimentos. Se normalizan los pesos para tener vectores unitarios, de ahí el denominador.

$$p_{ij} = \frac{(0{,}5 + 0{,}5\frac{f_{ij}}{\text{máx}_k(f_{ik})}) \cdot itf_j}{\sqrt{\sum_{u=1}^{m}(0{,}5 + 0{,}5\frac{f_{iu}}{\text{máx}_k(f_{ik})})^2 \cdot itf_u^2}} \tag{3}$$

donde

- f_{ij} es el número de veces que el término t_i aparece en el documento d_j,
- $max_k(f_{ik})$ es el máximo de los valores de frecuencia para el término t_i en toda la colección de documentos (es decir, será el valor f_{ik}, siendo d_k el documento en que más aparece),
- $itf_j = log\frac{n}{|d_j|}$ es el '*inverse term frecuency*' para el documento d_j, siendo $|d_j|$ el número de términos diferentes existentes en ese documento.

El cálculo del *inverse term frecuency* muestra que un documento corto juega un papel más importante que uno largo. Si dos términos coocurren en un documento largo, la probabilidad de que sean similares es menor que si coocurren en uno corto.

Una vez obtenido la representación de términos en función de los documentos, el paso siguiente es calcular la similitud entre dos términos, t_i y t_j. Para ello se emplea el producto escalar.

$$\text{SIM}(t_i, t_j) = \vec{t_i} * \vec{t_j}^T = \sum_{k=1}^{m} p_{ik} \cdot p_{jk} \tag{4}$$

El cálculo para todos los términos de la colección produce el tesauro de similitud. Se trata de una matriz simétrica, cuyos valores están también entre 0 y 1, igual que la matriz de asociación. Debemos resaltar que la construcción de la matriz de similitud es computacionalmente mucho más costosa que la matriz de asociación. Este es un aspecto importante a la hora de aplicar la expansión en sistemas reales.

2.3. Expansión de la consulta

Como ya hemos indicado el objetivo al utilizar una u otra matriz es expandir la consulta completa, y no solo cada término individual por separado. Para que un término pueda añadirse a la consulta debe tener una similitud alta entre dicho término y todos los términos de la consulta. En el modelo vectorial la consulta q se representa por un vector, $\vec{q} = (q_1, q_2, \ldots, q_n)$, donde q_i es el peso del término t_i en la consulta. Cada término de la consulta está relacionado con el resto de términos de la colección, según la matriz de asociación/similitud, Ec. (1). Para obtener los términos más similares con toda la consulta se expande cada término de la misma y se van sumando los valores de asociación/similitud correspondientes. Sea t un término de la colección, el valor de similitud con toda la consulta se obtiene aplicando la siguiente ecuación:

$$\text{sim}(q, t) = \text{sim}(\sum_{t_i \in q} q_i t_i, t) = \sum_{t_i \in q} q_i \cdot \text{REL}(t_i, t) \tag{5}$$

La función $\text{REL}(t_i, t)$ debe ser sustituida por la Ec. (2) ó (4), dependiendo de la matriz que se desea utilizar. En cualquier caso, para cada término de la colección se obtiene un valor de relación con toda la consulta. Pueden ordenarse los términos de acuerdo a ese valor, y expandir la consulta con aquellos que tengan los valores más altos. Notaremos por r el número de términos que se añadirán a la consulta original.

2.4. Pesado de términos expandidos

Falta por determinar el peso, en el espacio de términos, asociado a cada término t_e que se añadirá a la consulta. En [11] se utiliza la suma de los pesos de los términos iniciales para reducir el valor de similitud obtenido, Ec. (6). Los pesos de términos de la consulta original pueden llegar a modificarse si aparecen entre los r primeros seleccionados para expandir.

$$q_e = \frac{\mathrm{sim}(q, t_e)}{\sum_{t_i \in q} q_i} \tag{6}$$

El denominador de la Ec. (6) es un coeficiente que sirve para reducir el peso de los términos expandidos respecto de los términos originales. El objetivo es reducir su influencia en la consulta completa. Sin embargo, no es la única expresión posible que se puede aplicar. Incluso se puede aplicar un coeficiente unidad como valor asociado de similitud con toda la consulta para dichos términos, sin reducir en absoluto su valor. Nuestros experimentos han ido encaminados también a determinar la influencia de ese coeficiente en los resultados finales.

2.5. Normalización de la consulta original

En este punto es necesario que hagamos un comentario sobre la normalización de los valores de los pesos de la consulta original. En todo el razonamiento seguido hasta ahora no se ha hecho referencia a la normalización de los vectores que representan a documentos y consultas en el espacio vectorial de términos. La normalización tiene su importancia a la hora de determinar la similitud entre documentos y consultas. Para medir tal similitud se suele utilizar el producto escalar de los vectores que representan a documentos y consultas [19]. De este manera se pueden presentar al usuario los documentos ordenados por valor de similitud. Al realizar el producto escalar de dos vectores, por otra parte, se asegura que el valor de similitud se encuentra entre 0 y 1. Para evitar tener que calcular en cada momento el módulo de los vectores, éstos suelen normalizarse a priori.

La normalización para los vectores de los documentos tiene además otra razón. Esa normalización evita que documentos largos, en los cuales los términos aparecen con mucha frecuencia, prevalezcan sobre documentos cortos que pueden ser igualmente relevantes. La normalización en el vector consulta solamente provoca que la similitud se encuentre entre 0 y 1, pero nada más, pues el orden en que se mostrarían los documentos recuperados al usuario sería el mismo aunque no se realizase tal normalización.

Sin embargo, la normalización de la consulta original influye en el valor del denominador de la Ec. (6). Es decir, no sólo la elección del coeficiente para el pesado de términos expandidos es un aspecto que puede llegar a ser crítico, sino también la aplicación de normalización sobre la consulta original. En este trabajo también se presenta la influencia de estos aspectos en la expansión de consultas.

3. Experimentos llevado a cabo

En nuestros experimentos hemos empleado la colección de prueba utilizada en varios de nuestros trabajos [23] en el Taller CLEF [10]. Las características de la colección se muestran en el Cuadro 1. La colección de documentos proviene de la agencia de noticias española EFE, de todas las noticias del año 1994. Se trata de 215.718 documentos en español (513 MB). Hemos utilizado los campos `TITLE` y `TEXT` de los mismos. Los experimentos se han realizado con las consultas 41 a 90 de la colección en español, y se ha utilizado exclusivamente el campo `ES-title` de las mismas.

Cuadro 1: Características de la colección.

Colección	EFE'94
Nº de documentos	215.738
Nº de consultas	50
Nº de términos índice	352.777
Nº medio de palabras por documento	333,68 (máx. 2210,mín 9)
Nº medio términos índice únicos por documento	120,48
Nº medio términos índice únicos por consulta	2,73

En esta colección hemos considerado que un término índice está compuesto por caracteres alfanuméricos, sin considerar acentos y reduciendolos a minúscula. El número medio de términos se ha considerado después de eliminar las palabras vacías. No se ha aplicado lematización. Una vez obtenidos los términos índice, la representación de documentos y consultas se ha realizado utilizando el mecanismo de pesado *tf-idf* y las recomendaciones de [13]. En todos los experimentos hemos normalizado los vectores de los documentos, no así el de consultas, pues deseamos ver también su influencia en los resultados. Asimismo hemos empleado el producto escalar para el cálculo de similitudes entre documentos y consultas.

A partir de la colección documental hemos construido las matrices de asociación y similitud. Debido a la poca extensión de los documentos no se han utilizado ventanas de palabras o caracteres como suele ser habitual en otros experimentos. Las matrices obtenidas arrojan resultados comparables. Por ejemplo, en el Cuadro 2 se muestran los 20 términos mejor relacionados para la palabra *terremoto*.

En la expansión de cada consulta se ha calculado el tesauro local para cada término de la misma, de acuerdo a la Ec. 5 (aplicando el criterio de relación que corresponda), y se han ordenado los términos relacionados en orden decreciente para luego tomar los r primeros. Los

Cuadro 2: Ejemplo para la entrada *terremoto* en las matrices de expansión.

Asociación						**Similitud**	
Tanimoto		Coseno		Dice			
terremoto	1,0000	terremoto	1,0000	terremoto	1,0000	terremoto	1,0000
richter	0,4058	richter	0,5827	richter	0,5773	richter	0,6192
seismo	0,3502	seismo	0,5288	seismo	0,5188	seismo	0,5491
epicentro	0,2800	epicentro	0,4569	epicentro	0,4375	epicentro	0,4833
temblor	0,2045	temblor	0,3626	temblor	0,3395	escala	0,3993
escala	0,1855	escala	0,3488	escala	0,3130	grados	0,3716
grados	0,1844	grados	0,3289	grados	0,3113	temblor	0,3696
sacudio	0,1725	sacudio	0,3255	sacudio	0,2943	sacudio	0,3525
magnitud	0,1704	terremotos	0,3018	magnitud	0,2912	magnitud	0,3380
terremotos	0,1407	magnitud	0,2935	terremotos	0,2467	terremotos	0,3173
temblores	0,1205	sismico	0,2792	temblores	0,2151	temblores	0,2860
intensidad	0,1137	temblores	0,2721	intensidad	0,2041	sismico	0,2798
sismico	0,1080	seismos	0,2591	sismico	0,1949	seismos	0,2603
seismos	0,1022	sismica	0,2424	seismos	0,1854	sismica	0,2538
sismica	0,0929	daños	0,2130	sismica	0,1700	intensidad	0,2405
daños	0,0913	northridge	0,2126	daños	0,1673	northridge	0,2400
damnificados	0,0833	intensidad	0,2092	damnificados	0,1537	daños	0,2379
sacude	0,0737	tsunami	0,2056	sacude	0,1373	tsunami	0,2221
telurico	0,0729	maremoto	0,2026	telurico	0,1358	sismicos	0,2121
sintio	0,0706	sismicos	0,2006	sintio	0,1318	maremoto	0,2099
olas	0,0674	sacude	0,1879	olas	0,1263	sacude	0,2061

resultados se han evaluado considerando tres valores medios: precisión media para todas las consultas de la colección, R-precisión media, y precisión media a 20 documentos vistos. Puede ver una descripción de estas medidas en [23]. Hemos incluido la medida de la precisión media a 20 documentos vistos porque los sistemas de RI y motores de búsqueda en Internet normalmente presentan los resultados en grupos de 10 ó 20 documentos. Además, la mayoría de usuarios solamente muestra interés por los resultados de la primera pantalla [18].

3.1. Experimento 1

Hemos realizado tres tipos de experimentos. El primero de ellos trata de determinar si la expansión de consultas produce mejoría en los resultados del sistema de recuperación. Para llevarlo a cabo se han considerado los mismos criterios que en [11]: consultas iniciales sin normalizar y se ha aplicado la Ec. (6) para el pesado de términos expandidos. En el Cuadro 3 se muestran los resultados de las tres medidas para un valor de $r = 200$. Podemos ver que con cualquier tipo de expansión utilizando tesauros de asociación o de similitud se aprecia mejoría. Ya hemos indicado que la construcción de la matriz de similitud es computacionalmente más costosa que la matriz de asociación. Podemos ver que la medida para el Coseno es parecida a la de Similitud.

La Figura 1 muestra los resultados de precisión media, R-precisión media y precisión media a 20 documentos vistos, respectivamente, en función del número de términos expandidos. Podemos apreciar que la mejora sigue la misma evolución para las tres pruebas. Los mejores resultados se obtienen con las matrices Coseno y Similitud. Apreciamos una mejora importante con la expansión de unos pocos términos, pero a partir de cierto número (en torno a 200) la mejoría crece muy lentamente, y sigue creciendo, quizás indefinidamente. Esto parece coincidir con experimentos de otros autores [11] cuando la colección de pruebas es grande. Quizás esto tenga que ver con el hecho de que se necesitan más términos discriminantes en colecciones de prueba de gran tamaño.

3.2. Experimento 2

El segundo tipo de experimentos mide la influencia en la expansión de consultas del coeficiente aplicado al peso de los términos expandidos, en función del número de términos expandidos. Se consideran los coeficientes que se muestran en el Cuadro 4, en el que k y mod son el número de términos y el módulo de la consulta original, respectivamente. El primero de los coeficientes es la suma de los pesos de la consulta original, tal como se indica en la Ec. (6). El segundo considera solamente el número de términos de la consulta original. El coeficiente 'Mágico' es un

Cuadro 3: Resultados del primer experimento ($r = 200$).

		Tesauro Asociación			**Tesauro**
Medidas	**Original**	Tanimoto	Coseno	Dice	**Similitud**
Precisión media	0,3179	0,3296(3,68 %)	0,3414(7,39 %)	0,3342(5,13 %)	0,3403(7,05 %)
R-Precisión media	0,3249	0,3363(3,51 %)	0,3440(5,88 %)	0,3402(4,71 %)	0,3483(7,20 %)
Precisión a 20 docs	0,3184	0,3347(5,12 %)	0,3459(8,64 %)	0,3367(5,75 %)	0,3459(8,64 %)

Cuadro 4: Coeficientes para el pesado de términos expandidos.

Qiu-Frei	Media	Mágico	Unidad
$\frac{1}{\sum_{t_i \in q} q_i}$	$\frac{1}{k}$	$\frac{1}{mod * sqrt(k)}$	1

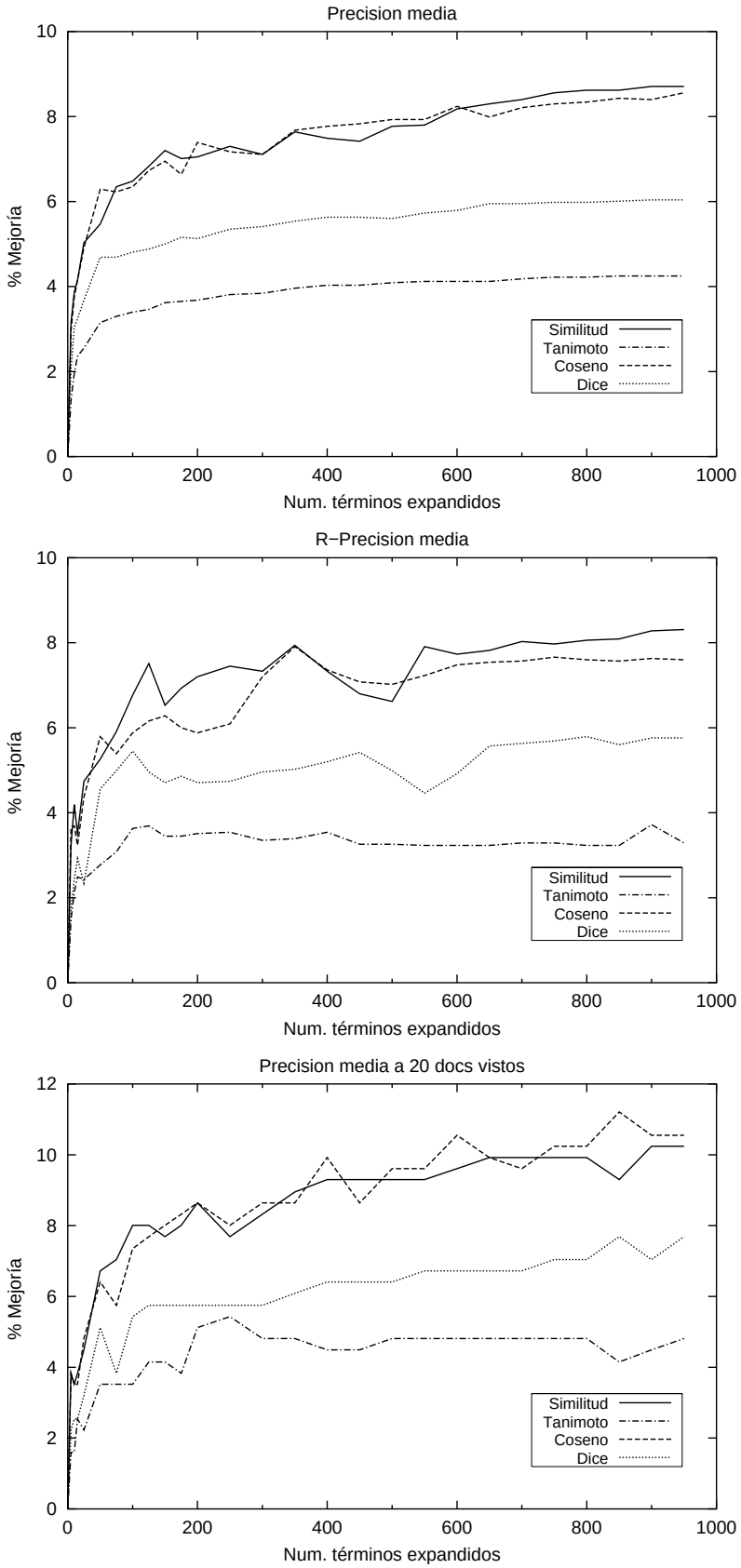

Figura 1: Experimento 1.

coeficiente artificial, cuya aplicación no tiene significado especial, pero que nos permiten valorar la influencia en la recuperación. Finalmente, el último coeficiente en la unidad, esto es, no se reduce la influencia de los términos añadidos respecto de los originales en la consulta expandida.

Computacionalmente el menos costoso de los coeficientes es el segundo, aparte, claro está, del último. Esto es importante a la hora de aplicar la expansión en entornos reales, con tiempos críticos de respuesta.

Para llevar a cabo este experimento hemos considerado solamente la expansión de consultas con tesauro de similitud, ya que, junto con el correspondiente al Coseno, presenta el mejor comportamiento. Además hemos considerado la situación de consultas iniciales sin normalizar. Para comprobar la evolución en función del número de términos expandidos, se ha tomado como medida la precisión media.

La Figura 2 muestra los resultados de este segundo experimento. Podemos apreciar que destaca el coeficiente 'Media' del resto. La prueba que consigue la peor evolución es la que hemos denominado 'Unidad': para unos pocos términos expandidos (hasta 50) los resultados empeoran hasta casi un 7 % respecto de las consultas originales, pero con el aumento del número de términos expandidos enseguida se consiguen valores de mejora similares a los otros coeficientes.

3.3. Experimento 3

El tercer tipo de experimentos que hemos realizado determina la influencia de la normalización de la consulta original en los resultados de expansión, en relación con el coeficiente aplicado. Se han realizado las pruebas utilizando todos los coeficientes del Cuadro 4. Una normalización en la consulta inicial afecta a los coeficientes 'Qiu-Frei' y 'Mágico', pues en ellos interviene el peso de los términos iniciales de la consulta, y éstos son diferentes con consultas normalizadas y sin normalizar. Sin embargo, los coeficientes 'Media' y 'Unidad' no se ven afectados.

El objetivo de este experimento es comparar los resultados con el experimento anterior, y analizar la influencia de la normalización inicial. De nuevo, esto es importante a la hora de aplicar la expansión en sistemas de recuperación reales, ya que una normalización de las consultas iniciales conlleva un gasto computacional, que podría evitarse con la aplicación de coeficientes aceptables comparando ambas situaciones.

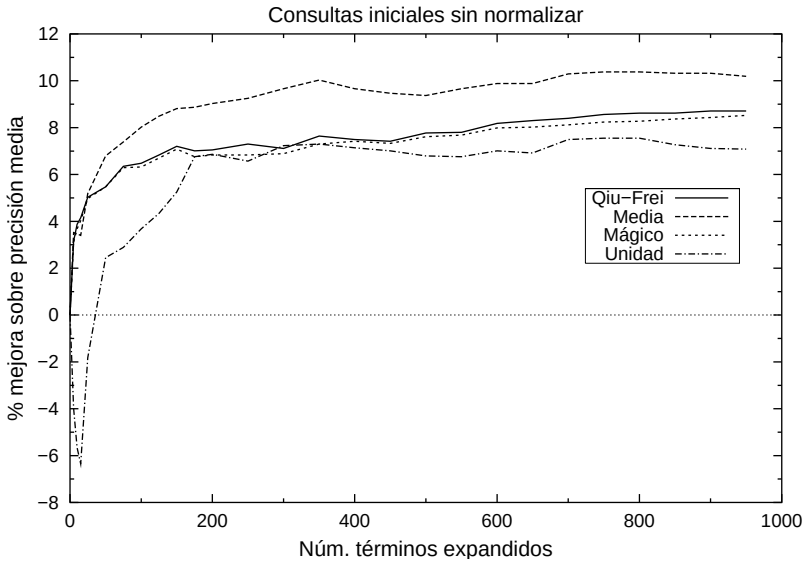

Figura 2: Experimento 2.

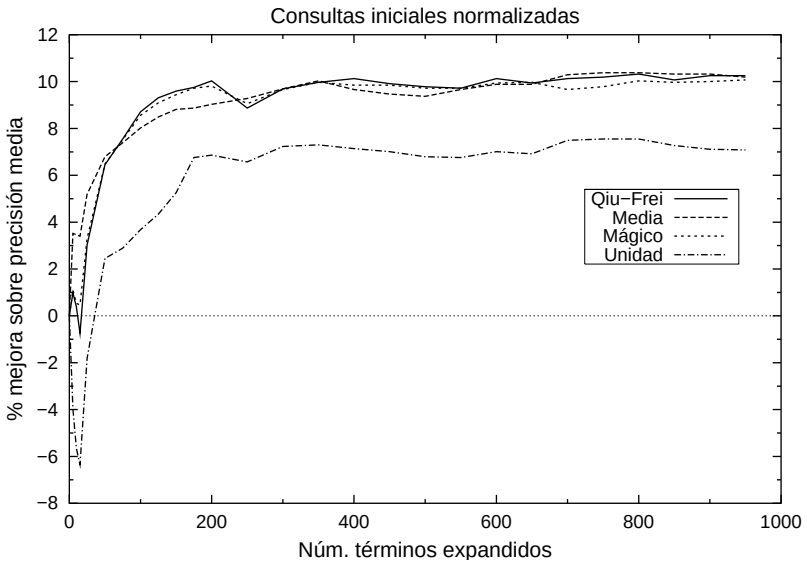

Figura 3: Experimento 3.

La Figura 3 muestra los resultados para este experimento. Para un número elevado de términos añadidos a la consulta se aprecian unos resultados ligeramente superiores a los del experimento anterior. Ahora bien, la diferencia se aprecia sobre todo cuando se expande la consulta original con unos pocos términos, por debajo de 25. En esta situación los resultados de recuperación son peores incluso que para la consulta sin expandir.

Parece aconsejable, entonces, utilizar en cualquiera de los casos la prueba que hemos denominado 'Media' pues es la que obtiene mejores resultados. Debemos señalar que al utilizar este coeficiente los resultados son independiente de que la consulta inicial esté o no normalizada, mientras que el resto de pruebas es dependiente de esta circunstancia. Por otra lado, computacionalmente es un mecanismo que consume poco tiempo.

4. Conclusiones

Se ha descrito una técnica que obtiene buenos resultados en expansión de consultas. Se han construido dos tesauros de forma automática a partir de la colección documental sobre la que se lanzan las consultas, uno de asociación y otro de similitud. La matriz de asociación se ha construido utilizando medidas simples de coocurrencia de términos. La de similitud se ha realizado aprovechando las posibilidades que poseen los documentos para representar a los términos. En ambos casos, una vez se tiene construida la matriz, para una consulta dada se obtiene un conjunto de términos que poseen un alto grado de relación con toda la consulta, y no sólo con cada término individual por separado. La elección y peso de los términos expandidos se ha realizado considerando este hecho. Estos dos aspectos son los más importantes a tener en cuenta en la expansión de consultas. Los experimentos llevados a cabo sobre la colección CLEF muestran una mejoría en los resultados de recuperación.

Además, en contraposición a otras técnicas de análisis local que basan su expansión en función de los términos de unos pocos documentos con criterios de relevancia por parte del usuario, aquí se ha utilizado una técnica global que relaciona entre sí todos los términos de la colección de pruebas. Efectivamente, los resultados son algo peores, motivado fundamentalmente porque la expansión que hemos utilizado no incorpora ningún tipo de información adicional de relevancia, que sí incorporan esos métodos. Se trata por tanto de un mecanismo que puede

ejecutarse automáticamente como una característica adicional más del sistema de recuperación de información. El inconveniente fundamental está en el coste computacional que requiere la construcción de los tesauros. La tarea es más sencilla para documentos nuevos que se incorporen a la colección documental, pues únicamente se necesita modificar las relaciones de los términos que aparezcan en esos nuevos documentos.

En esta situación se ha analizado el mecanismo de expansión para determinar la dependencia de dos aspectos en los resultados finales. El primero de ellos es la aplicación de un coeficiente reductor para el valor del peso de los términos expandidos. Ese coeficiente no es necesario cuando el número de términos expandidos es elevado, ya que, según las pruebas realizadas, siempre hay mejoría de resultados. Sin embargo, es necesario aplicar algún coeficiente reductor cuando el número de términos expandidos es pequeño, menor de 50, para evitar empeorar los resultados. Parece aconsejable utilizar cualquiera de los métodos 'Qiu-Frei' o 'Media'.

El segundo aspecto que influye en el mecanismo de expansión es la normalización de la consulta original. Hemos repetido las pruebas utilizando los mismos coeficientes, pero para consultas iniciales normalizadas. Para un número pequeño de términos expandidos, menos de 25, es muy importante elegir correctamente dicho coeficiente. Para un número de términos mayor todos los coeficientes ofrecen más o menos la misma mejoría.

Un aspecto que no debe dejarse de lado es el coste computacional de la expansión de consultas. El hecho de incluir más términos a la consulta provoca un empeoramiento en el tiempo de respuesta del sistema. Por eso es importante reducir todo lo posible el resto de tiempos. En primer lugar es conveniente no realizar normalización de la consulta original que suponga un tiempo adicional. En segundo lugar es necesario buscar coeficientes de aplicación rápida, que permitan asimismo mejorar los resultados.

Para nuestra colección documental hemos encontrado un coeficiente rápido e independiente de la normalización de la consulta original, que obtiene unos porcentajes de mejoría elevados. Es el coeficiente que hemos denominado 'Media'. No obstante, faltaría realizar los experimentos con otras colecciones documentales para poder recomendar su uso de forma general.

Por otra parte, se han utilizado consultas con muy pocos términos. Considerando que las consultas de los motores de búsqueda en Internet típicamente suelen contener uno, dos o tres términos, esta técnica puede ser especialmente útil en ese contexto. Ahora bien, considerando que la volatilidad de la información en Internet es muy elevada, y que la modificación del tesauro sería importante, pues desaparecerían y se incorporarían una cantidad enorme de documentos a la colección, creemos que esta técnica únicamente es utilizable en situaciones mucho más estáticas.

Referencias

[1] H. Billhardt, D. Borrajo, and V. Maojo. A context vector model for information retrieval. *Journal of the American Society for Information Science and Technology*, 53(3):236–249, 2002.

[2] C. J. Crouch and B. Yang. Experiments in automatic statistical thesaurus construction. In N. J. Belkin, P. Ingwersen, and A. M. Pejtersen, editors, *Proceedings of the 15th Annual International ACM SIGIR Conferenceon Research and Development in Information Retrieval. Copenhagen, Denmark, June 21-24, 1992*, pages 77–88. ACM, 1992.

[3] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Comunications of the ACM*, 30(11):964–971, November 1987.

[4] G. Grefenstette. Use of syntactic context to produce term association lists for text retrieval. In N. J. Belkin, P. Ingwersen, and A. M. Pejtersen, editors, *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, Denmark, June 21-24, 1992*, pages 89–97. ACM, 1992.

[5] C. Han, H. Fujii, and W. Croft. Automatic query expansion for japanese text retrieval. Technical Report UM-CS-1995-011, Department of Computer Science, Lederle Graduate Research Center, University of Massachusetts, 1995. On line:`ftp://ftp.cs.umass.edu/pub/techrept/techreport/1995/UM-CS-1995-011%.ps`.

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

[7] Y. Jing and W. B. Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO-94, 4th International Conference "Recherche d'Information Assistee par Ordinateur"*, pages 146–160, New York, US, 1994.

[8] J. Minker, G. G.A. Wilson, and B. Zimmerman. An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval*, 8(6):329–348, 1972.

[9] H. J. Peat and P. Willet. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5):378–383, 1991.

[10] C. Peters and M. Braschler. Cross-language system evaluation: The CLEF campaigns. *Journal of the American Society for Information Science and Technology*, 53(12):1067–1072, 2001.

[11] Y. Qiu and H.-P. Frei. Concept-based query expansion. In *Proceedings of the Sixteenth International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh (US), 1993.

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

[13] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[14] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New-York, 1983.

[15] G. Salton and C. S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4):351–372, 1973.

[16] H. Schutze. Dimensions of meaning. In *Proceedings of Supercomputing '92, Minneapolis, 1992*, pages 787–796, 1992.

[17] A. Smeaton and C. van Rijsbergen. The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 26(3):239–246, 1983.

[18] A. Spink, D. Wolfram, B. J. Jansen, and T. Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, 2001.

[19] C. van Rijsbergen. *Information Retrieval*. Dept. of Computer Science, University of Glasgow, second edition, 1979. Also on-line: `http://www.dcs.gla.ac.uk/Keith/`.

[20] E. Voorhees. Query expansion ussing lexical-semantic relations. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 61–69, 1994.

[21] D. Wolfram, A. Spink, B. J. Janses, and T. Saracevic. Vox populi: The public searching of the web. *Journal of the American Society for Information Science and Technology*, 52(12):1073–1074, 2001.

[22] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, 18(1):79–112, 2000.

[23] Á. F. Zazo, C. G. Figuerola, J. L. A. Berrocal, and R. Gómez. Recuperación de información utilizando el modelo vectorial. participación en el taller CLEF-2001. Technical Report DPTOIA-IT-2002-006, Departamento de Informática y Automática - Universidad de Salamanca, Mayo 2002. On line: `http://tejo.usal.es/inftec/2002/DPTOIA-IT-2002-006.pdf`.