Procesamiento del Lenguaje Natural, nº 33, septiembre de 2004 ISSN 1135-5948

SEPLN

Sociedad Española para el Procesamiento del Lenguaje Natural

Artículos

Procesamiento del Lenguaje Natural, núm. 33 (2004), pp. 1-8 recibido 31-3-2004; aceptado 10-5-2004

Incidencia de técnicas de realimentación de consultas en Recuperación de Información Distribuida

Fernando Martínez Santiago **Miguel Ángel García**
Juan Gutierrez Marín **L. Alfonso Ureña López**
Grupo Sistemas Inteligentes de Acceso a la Información
Departamento de Informática
Universidad de Jaén
e-mail:{dofer,magc,juangu,laurena}@ujaen.es

Resumen: Este artículo presenta un exhaustivo análisis sobre la aplicación de la realimentación por pseudo-relevancia (PRF) aplicada a la recuperación de información distribuida (DIR). Este trabajo se centra en la efectividad de la realimentación aplicada al problema de la fusión de colecciones. Normalmente, los sistemas distribuidos usan la realimentación de la misma forma que los sistemas tradicionales de recuperación de información. Para cada colección, los resultados mejoran usando PRF. Los sistemas DIR fusionan listas de documentos, cada una obtenida a partir de una colección previamente seleccionada para una consulta dada. En el momento en que tenemos un ranking de documentos global, podemos emplear esta lista para aplicar PRF de nuevo, ahora a nivel global mejor que a nivel local. Para aplicar PRF global, hemos desarrollado una nueva técnica llamada RSV en dos pasos (2-step RSV), asimismo describimos un número de experimentos involucrando los dos niveles —local y global— de aplicación de las técnicas de PRF.
Palabras clave: Fusión de Colecciones, DIR, TREC, CORI, 2-step RSV

Abstract: This paper presents a thorough analysis of the capabilities of the pseudo-relevance feedback (PRF) technique applied to Distributed Information Retrieval (DIR). This work emphasizes the effectiveness of PRF applied to the collection fusion problem. Usually, DIR Systems apply PRF in the same way of traditional Information Retrieval systems. For each collection, local results are improved through PRF. DIR systems merge the documents of rankings that are returned from a set of collections. Since a new global list of documents is available, we could use that list to apply PRF again, but at global level better than at local one. In order to apply global PRF, we have developed a new merging approach called 2-step RSV. Finally, we describe a number of experiments involving the two levels, local and global, of application of the PRF techniques.
Keywords: Collection Fusion, DIR, TREC, CORI, 2-step RSV

1. *Introducción*

Normalmente, un sistema DIR debe clasificar las colecciones de documentos por relevancia en base a una consulta, seleccionando el mejor conjunto de colecciones de un ranking, y fusionando la clasificación de documentos devuelta para cada una de las colecciones seleccionadas. Este último problema se conoce como el problema de la fusión de colecciones (Voorhees, Gupta, y Jhonson-Laird, 1995). El principal objetivo de este trabajo es un exhaustivo análisis de la realimentación por pseudo-relevancia (PRF), aplicada al problema de la fusión de colecciones.

La realimentación por pseudo-relevancia (J. J. Rocchio, 1971), es un proceso que mejora el rendimiento de la recuperación de información (RI). Su objetivo es recuperar y clasificar en los puestos más altos a aquellos documentos que son similares a los seleccionados como relevantes por el usuario. Por otro lado, la PRF no necesita interacción por parte del usuario, pero hace la suposición de que los N primeros documentos clasificados son relevantes. Así, esta técnica es más adecuada en un sistema real y transparente al usuario. La realimentación ha sido aplicada tanto en escenarios no distribuidos como en sistemas DIR, pero necesita la colaboración del usuario para decidir qué documentos son relevantes. Tradicionalmente los sistemas DIR aplican PRF de la misma forma que los sistemas de recuperación de información no distribuidos, a

ISSN 1135-5948

saber, la PRF o expansión de consultas se aplica en un entorno local por cada sistema individual de recuperación de información.

En este trabajo, usaremos realimentación local para referirnos a esta última manera de aplicar realimentación. Nuestro objetivo es aplicar PRF globalmente sobre la clasificación (el ranking o lista de documentos) final una vez fusionada. A este tipo de realimentación la denominamos PRF global. Trabajos anteriores en este área se han centrado en usar PRF como una vía de mejora del proceso de selección de colecciones, obteniendo unos resultados algo pobres (Ogilve y Callan, 2001). Nuestro trabajo explora PRF global como una manera de mejorar el proceso de fusión de documentos, no el proceso de selección de colecciones. Los documentos devueltos por cada motor de búsqueda se fusionan utilizando un algoritmo llamado RSV en dos pasos (2-step RSV) (Martínez-Santiago, Martín, y Ureña, 2003). Este algoritmo ha demostrado funcionar bien en sistemas CLIR basados en traducción de consultas, aunque su aplicación a entornos DIR requiere un esfuerzo adicional: aprender características de la colección tales como la frecuencia documental, tamaño de la colección, etc. Por otro lado, es posible aplicar PRF a nivel global mejor que a nivel local, ya que el 2-step RSV desarrolla un nuevo índice global basado en los términos de la consulta y el total de los documentos recuperados. Destacamos que el algoritmo RSV en dos pasos no implementa un sistema DIR completo, sino que se centra únicamente en la fusión de documentos. Así tanto el ranking como la selección de colecciones se realizan mediante el conocido algoritmo CORI (Callan, Lu, y Croft, 1995).

Cuadro 1: Sistemas y algoritmos DIR implementados para los experimentos

	CORI	2-step RSV
ranking de colecciones	CORI	CORI
selección de colecciones	CORI	CORI
fusión de documentos	CORI	2-step RSV
disponible PRF local	Yes	Yes
disponible PRF global	No	Yes

2. *Cálculo de la relevancia documental en dos pasos*

El cálculo de la relevancia documental en dos pasos o 2-step RSV consiste en agrupar las frecuencias documentales de un término daod de una consulta(Martínez-Santiago et al., 2003). El método requiere calcular la puntuación obtenida por cada documento cambiando la frecuencia documental de cada término que aparece en la consulta: dado un término de la consulta, su nueva frecuencia documental será el resultado de sumar a su frecuencia documental original la frecuencia documental alcanzada por tal término en el resto de las subcolecciones seleccionadas. Por ejemplo, si las colecciones I_1 e I_2 son seleccionadas, y la consulta contiene el término "goverment", entonces la nueva frecuencia global será $df_{I_1}(government) + df_{I_2}(government)$.

Dada una consulta los dos pasos son:

1. La fase de preselección de documentos se corresponde con el lanzamiento de la consulta sobre cada subcolección seleccionada I_j. Como resultado de unir los documentos más relevantes recuperados para cada colección obtenemos una única colección (lista) de documentos preseleccionados I'.
2. La fase de reordenamiento consiste en reindexar la colección (ahora ya será un ranking) I', pero considerando tan sólo el vocabulario de la consulta. Finalmente, se elabora un nueva consulta formada por los términos anotados y se lanza tal consulta sobre el nuevo índice.

La hipótesis de este método es como sigue: Las puntuaciones logradas por documentos en sistemas de IR autónomos no son comparables aun utilizando el mismo modelo IR, debido a que la puntuación alcanzada por el documento es relativa a la colección a la que pertenece. Es más, la frecuencia documental alcanzada por cada término (depende de la colección) es determinante en la puntuación alcanzada por cada documento. Es posible recalcular el peso de cada documento como si todas las colecciones independientes formaran parte de una única colección local, considerando las siguientes simplificaciones: a) sólo es necesario reindexar los términos aparecidos en la consulta inicial y b) sólo es necesario reindexar las listas de documentos devueltos por cada sistema IR independientemente. Para ello, sólo es necesario almacenar localmente una lista de términos índice junto con la frecuencia documental de tal término en cada colección considerada. Así, el sistema DIR debe enviar la consulta a cada colec-

ción seleccionada, descargas los documentos más relevantes, creas un nuevo índice con tales documentos por medio de una fórmula de pesado y, finalmente la consulta se formula contra el nuevo índice. Los documentos I' se ordenan utilizando un nuevo índice. Nótese que la creación de este índice en tiempo de consulta es posible porque el vocabulario del nuevo índice está compuesto sólamente por términos de la consulta, y porque el índice puede crearse sobre unos pocos documentos, conforme el usuario solicite más resultados, tal como se expone en la sección 2.2.

2.1. Elementos para el cálculo del índice global

En este tabajo, hemos adoptado la función OKAPI-BM25 (Robertson, Walker., y Beaulieu, 2000), como esquema de pesado durante la segunda fase, el cual conjuga tanto el peso dado a cada término de la consulta, como la expansión de ésta. OKAPI-BM-25 se formula como sigue:

$$\sum_{T \in Q} W^{(1)} \frac{(k_1+1)tf}{K+tf} qtf \qquad (1)$$

donde

- Q es una consulta conformada por los términos T.
- K se calcula como $K = k_1((1-b) + b * dl/avgdl)$.
- k_1 y b son constantes que se fijan en $k1 = 1,25$ y $b = 0,75$.
- tf es la frecuencia del término en un documento dado.
- qtf es la frecuencia del término en la consulta Q.
- dl y $avdl$ son respectivamente la longitud del documento y la longitud media del documento en la colección.
- $W^{(1)}$ es el peso del término T en Q propuesto por Robertson y Spark-Jones (Robertson y Jones, 1976).

$$W^{(1)} = \log \frac{(r+0,5)/(R-r+0,5)}{(n-r+0,5)/(N-n-R-r+0,5)} \qquad (2)$$

  - N es el número de documentos en la colección.
  - n es la frecuencia documental (número de documentos que contienen el término).
  - R es el número de documentos conocidos que son relevantes para la consulta dada.
  - r es el número de documentos relevantes que contienen el término.

La cuestión es cómo aplicar la fórmula globalmente (sobre la colección I') mejor que localmente (para cada colección I_i). Se conocen varios elementos y otros se aprenden o aproximan:

- La frecuencia del término se conoce porque los documentos deben descargarse por el sistema DIR antes del segundo paso del *2-step RSV*.
- La frecuencia del documento n es la suma de las frecuencias locales de los documentos n_i. El número de documentos devueltos es *df* del término. Además, existen varios enfoques para aprender este factor, tales como el conocido algoritmo *Query-based sample* (Callan, Connell, y Du, 1999).
- El número de documentos en la colección N se obtiene de la misma manera que la frecuencia documental global: agrupando cada una por tamaño colección local.; si el tamaño de cada colección no está disponible, se aproxima usando los algoritmos *Capture-recapture* (Liu et al., 2001) o *sample-resample*(Si y Callan, 2003b).
- Finalmente, la longitud del documento medio se aprende por medio del algoritmo *Query-based*.

2.2. Un posible sistema DIR basado en el cálculo 2-step RSV

De todos los factores involucrados en la fórmula 1, la frecuencia documental es el más costoso, ya que requiere descargarse cada documento como paso previo a evaluarlo.

Por ejemplo, si deseamos crear un sistema que muestre los documentos ordenados de 10 en 10 el procedimiento que minimiza la descarga, queda como sigue:

1. Sea la colección distribuida formada por L subcolecciones, $I = I_1, ...I_L$. Dada una

consulta, con L' ($L' \leq L$) el número de subcolecciones seleccionados por el sistema DIR (por ejemplo utilizando el algoritmo CORI).

2. El sistema descarga m documentos para cada subcolección $I_i, 1 \leq i \leq L'$. De esta manera, el sistema DIR descarga un conjunto de documentos $I' = \{I'_1, \cdots, I'_{L'}\}$, $|I'| = L' * m \geq 10$, y $I'_i = \{d_1^i \cdots d_n^i\}$, $|I'_i| = m$ es el conjunto de documentos descargados de I_i.
3. Los documentos I' son repesados y alineados de acuerdo al 2-step RSV.
4. Los 10 primeros documentos dentro las lista puntuada I' se eliminan de cada conjunto de documentos I'_i de donde han sido extraidos. Si hay algún subconjunto vacío I'_i (alguna lista local vacía) antes de llegar al número de resultados globales deseado, se descargan los m documentos de I_i y repesa el conjunto de nuevos documentos I'_i. Se repite el paso previo hasta que cada subconjunto I'_i contenga al menos un documento.
5. Mostrar los 10 primeros documentos al usuario.
6. Si el usuario requiere más documentos, volver al paso 2.

Si el sistema muestra H documentos para cada petición de usuario y los H primeros documentos están dentro de la misma colección I_i (este escenario es el peor caso), el monitor DIR necesita descargar $L' * m + H - m$.

Por ejemplo, dado $H = 10$, $L' = 10$,$m = 2$, el monitor DIR necesita descargar 28 documentos en el peor de los casos posibles. Antes el sistema muestra los 10 primeros documentos. Nótese que los 28 documentos son los casos para los 10 primeros documentos, seguido de la petición de usuario que necesita descargar al menos 18 documentos

Por otra parte, algunos sistemas permiten descargar varios documentos a la vez en lugar de uno a la vez. Si esta característica no está disponible, aún es posible aplicar procesamiento paralelo para disminuir el tiempo de descarga necesario.

En este punto la aplicación de la realimentación mixta es fácil. Dados los R primeros documentos a nivel global, se aplica la fórmula 2. En este trabajo se analizan los 10 primeros documentos. Entonces se aplica la expansión de consulta[1] repesando cada documento descargado.

3. *Experimentos y Resultados*

3.1. Metodología de Experimentación

Los pasos seguidos para realizar cada experimento son los siguientes[2]:

1. Generar para cada colección un índice (OKAPI).
2. Se aplican algoritmos de muestreo basados en la consulta para obtener el modelo del lenguaje: vocabulario, tamaño medio del documento y tamaño de la colecciones. La frecuencia de los términos del documento, del vocabulario aprendido, se devuelve, para cada colección local, utilizando consultas con un único término. La suma de los documentos devueltos para cada colección es justo la frecuencia de documento para cada consulta.

Una vez que los índices están creados y el modelo de lenguaje aprendido, el sistema se evalúa. Para evaluar el método propuesto hemos usado los conjuntos de consultas de las conferencias TREC1 y TREC2, un total de 100 consultas. En todos los experimentos se han utilizado los campos título y descripción excluxivamente. Para cada consulta:

1. Utilizamos el algoritmo CORI para puntuar cada colección para esa consulta.
2. Elegimos las más prometedoras mediantes un algoritmo de *clustering*, como el descrito en (Callan, Lu, y Croft, 1995).
3. Cada colección seleccionada obtiene una lista local de documentos lanzando de nuevo cada consulta sobre el índice local.
4. En ciertos experimentos, aplicar realimentación por pseudo-relevancia (PRF) en cada colección, a nivel local. Se ha utilizado OKAPI-BM25 (eq. 1 y 2). Se tienen en cuenta los 10 primeros términos de los 10 primeros documentos.
5. La fusión de documentos se ha realizado utilizando los algoritmos de fusión CORI y 2-step RSV.

[1] Nótese que la expansión de la consulta se aplica sólo a nivel global y no localmente.

[2] El escenario tiene un diseño off-line.

6. En ciertos experimentos volver a aplicar PRF a nivel global sobre el índice on-line creado por 2-step RSV. La configuración de la realimentación a nivel global es similar que la que se aplica a nivel local.

Finalmente, la evaluación se ha medido en términos de precisión y cobertura. La medida de precisión es la precisión media obtenida en los 5,10,20 y 100 primeros documentos, así como la precisión media en 11 puntos de cobertura.

3.2. Descripción de las colecciones de prueba

Los experimentos se han realizado utilizando tres particiones de las conferencias TREC1 y TREC2. Las colecciones corresponden con textos publicados entre 1987 y 1990 en diversos diarios, agencias de noticias y editoriales. En total, son más de dos gigabytes de datos repartidos entre 740.000 documentos.

Sobre estas trece colecciones se han realizado tres juegos de prueba, descritos en la tabla 2:

- **TREC-1**. Las 13 colecciones indexadas bajo un único índice. Representa el mejor caso posible.
- **TREC-13**. Cada una de las 13 colecciones originales han sido indexadas separadamente.
- **TREC-80**. Las 13 colecciones originales se ha particionado hasta crear 80 subcolecciones, cada una de las ellas indexadas separadamente. Estas ochenta colecciones se han creado por origen y mediante una distribución aleatoria.

Cuadro 2: Descripción de las colecciones de prueba.

	# of docs.			Size (in MB)		
	Min.	Med.	Max.	Min.	Med.	Max.
TREC-1	741.991	741.991	741.991	2168	2168	2168
TREC-13	10.163	57.066	226.087	33	159,23	260
TREC-80	2473	9273	32.401	23	25,81	30

3.3. Experimentos sin expansión de consultas

En esta sección se compara CORI y 2-step RSV. No se ha aplicado expansión de consultas ni local y globalmente.

La tabla 3 muestra los resultados obtenidos sin realimentación. 2-step RSV mejora a CORI con el conjunto de colecciones TREC-13 (38.2 %) y con el TREC-80 (12.8 %). Estos resultados muestran que el incremento del rendimiento de 2-step RSV sobre CORI es dependiente de la colección, dado que la diferencia de mejora entre los dos conjuntos de colecciones está entorno al 25 %. Muchos trabajos muestran que el rendimiento de CORI es dependiente de la colección (Si y Callan, 2003a; Si y Callan, 2003b) mientras que 2-step RSV obtiene un rendimiento más estable. Además, la diferencia de mejora podría ser debida a los rendimientos diferentes de ambos algoritmos.

3.4. Experimentos con expansión de consultas

En esta sección se estudia la incidencia del uso de técnicas de expansión de consultas basadas en PRF local y global:

- Realimentación por pseudo-relevancia local. Cada colección local aplica localmente PRF con la finalidad de mejorar los resultados locales obtenidos.
- Realimentación por pseudo-relevancia global. Esta variante tan sólo es posible en el caso que se aplique 2-step RSV. Ya que el sistema DIR genera un nuevo índice global, es posible aplicar PRF sobre ese índice.
- Realimentación por pseudo-relevancia local y global. Finalmente, es posible aplicar PRF para cada colección local en un primer momento, y también para el sistema DIR después de forma global.

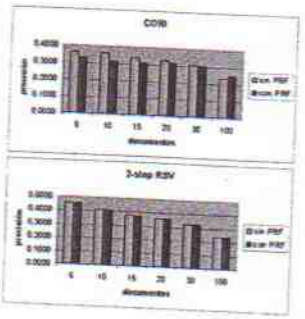

Figura 1: Impacto de la realimentación local

3.4.1. Experimentos con PRF Local

Tal como muestra la figura 1, la realimentación local no aporta mejora en el caso

Cuadro 3: Experimentos DIR sin realimentación (juego TREC-13, consultas 101-150)

Fusion	5-prec	10-prec	20-prec	100-prec	Avg-prec
CORI	0.468	0.416	0.363	0.265	0.130
2-step RSV	0.480	0.458	0.426	0.329	0.181
Centralizado	*0.492*	*0.492*	*0.444*	*0.346*	*0.194*

de 2-step RSV. Si se aplica CORI la situación mejora ligerísimamente sobre 13 colecciones. Sobre 80 colecciones, el uso de PRF incluso empeora algo los resultados, y no siempre. En cualquier caso las variaciones son tan pequeñas (no más de dos puntos) que la única conclusión es que PRF no afecta al resultado final, tanto para CORI como para 2-step RSV. En sistemas que no son completamente colaborativos este resultado era ya conocido para el caso de CORI. Para corroborar esta impresión, hemos aplicado el test de Wilcoxon sobre las precisiones medias alcanzadas en experimentos con la colección TREC-80, índices mixtos y OKAPI, y diversos algoritmos de selección de colecciones (*clustering*, *top-5*, *top-10* y *top-20*). Entre todos, son 16 experimentos destinados a comparar 2-step RSV con y sin realimentación local. El resultado obtenido arroja un p-valor igual a 1, con un nivel de error del 5 %, por lo cual podemos afirmar que no es posible encontrar diferencias significativas entre los resultados con y sin realimentación local, cuando se aplica 2-step RSV .

En el caso de CORI, de nuevo se ha recurrido a la prueba de Wilcoxon, aplicándolo sobre la misma población que en el caso anterior, pero tomando ahora las precisiones medias alcanzadas mediante el uso exclusivo de CORI. En este caso el p-valor alcanzado es de 0,0034. Este valor es lo suficientemente bajo como para rechazar la hipótesis de partida, por lo que podemos concluir que en el caso de CORI sí que se aprecia una mejora con el uso de PRF. El problema es que esa mejora es tan pequeña[3] que, desde el punto de vista de la recuperación de información, es despreciable, aunque la mejora no sea producto del azar.

Por todo ello, la única conclusión que es posible aventurar es que PRF local no afecta al resultado final o lo hace en una proporción despreciable.

(Ogilve y Callan, 2001) demuestran que el uso de técnicas de expansión de consultas (en su caso, *Local Context Analysis*) no mejoran ni la selección de colecciones ni apenas la de documentos. Como posibles causas de este comportamiento apuntan la longitud de la consulta una vez expandida, pues su nuevo tamaño dificulta la normalización de la puntuación alcanzada. Este razonamiento no es aplicable a 2-step RSV. Experimentos reportados en la sección anterior muestran que el rendimiento de 2-step RSV permanece invariable con el tamaño de la consulta. Posiblemente en el caso de 2-step RSV el motivo sea doble:

- El sistema DIR sólo maneja el vocabulario de la consulta original, por lo que aquellos documentos que hayan resultado ser relevantes en virtud de la expansión de consultas, pasarán desapercibidos.
- 2-step RSV no utiliza la puntuación alcanzada por cada documento localmente. La única condición de relevancia de un documento para 2-step RSV es su pertenencia a la lista devuelta por la colección local y el vocabulario de la consulta que tal documento contiene, nunca la puntuación alcanzada localmente.

3.4.2. Experimentos con PRF global

Tanto si se utiliza realimentación localmente como si no, es posible aplicar técnicas de expansión de consultas no considerando cada colección por separado, sino el índice creado en la segunda fase del método 2-step RSV. El costo computacional en este caso no se mucho más elevado que el que tendría en un sistema centralizado, pues sólo requiere analizar unos pocos documentos, en nuestros experimentos los 10 primeros documentos, con lo que en general es suficiente descargar tan sólo dos o tres documentos por colección seleccionada, siguiendo algún procedimiento como el descrito en la sección 2.

Un ejemplo de los resultados obtenidos se muestran en la tabla 4.

La mejora introducida en términos de precisión media con el uso de PRF global respec-

[3] En el caso de CORI, la mayor diferencia de precisión cuando se usa o no PRF es de tan sólo 0,0049 puntos

Cuadro 4: Experimentos DIR con realimentación global (juego TREC-80)

Fusión	5-prec	10-prec	20-prec	100-prec	Avg-prec
clusterig					
CORI	0.368	0.362	0.328	0.234	0.079
2-step RSV	0.456	0.412	0.362	0.241	0.089
2-step RSV+PRF global	0.440	0.408	0.383	0.274	0.105
Centralizado	*0.492*	*0.492*	*0.444*	*0.346*	*0.194*
Centralizado+PRF	*0.540*	*0.526*	*0.497*	*0.418*	*0.273*

to a 2-step RSV original es bastante significativa situándose en torno al 20 %.

La ganancia alcanzada por el uso de PRF en el modelo centralizado es de un 41 %, lo cual es bastante más del 20 % registrado por 2-step RSV. El motivo es claro, el modelo centralizado tiene acceso a todos los documentos de todas las colecciones, lo que posibilita la inclusión de documentos relevantes no seleccionados previamente. Esto, sin embargo, es imposible en el caso de 2-step RSV, ya que éste sólo considera *algunos* documentos y *algunas* colecciones, limitándose a reordenar los documentos ya seleccionados, pero nunca añadiendo documentos nuevos. En cualquier caso, esta situación podría cambiar sin más que lanzar la consulta una vez expandida sobre cada colección seleccionada, y aplicando sobre los nuevos resultados el algoritmo 2-step RSV original.

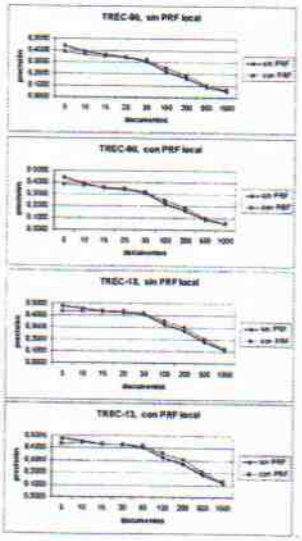

Figura 2: Impacto de la realimentación global

Si bien parece claro que PRF global mejora la precisión media, no necesariamente mejora le selección de los primeros documentos, siendo muy frecuente registrar peores precisiones con PRF global que sin PRF cuando se consideran tan sólo los cinco o diez primeros documentos. Este efecto es mostrado gráficamente en la figura 2. Se aprecia que conforme se aumenta el número de documentos mejora la precisión obtenida con PRF global, para finalmente obtener una precisión media y una R-precision sensiblemente superior (ver figura 3). El uso de PRF global mejora en general la cobertura, pues consigue introducir más documentos relevantes entre los mil primeros, pero no mejora la precisión desde los primeros documentos. ¿Es siempre pues aconsejable utilizar PRF global?. Como es usual en estos casos, la respuesta depende de las necesidades del usuario. En general aplicar PRF va a empeorar muy ligeramente el ranking de los 10 ó 15 primeros documentos, para a partir de ahí mejorar. Por otra parte, el coste computacional de aplicar PRF es moderado pero no nulo, pues requiere analizar los primeros documentos en tiempo de consulta. Posiblemente sea este coste computacional el que realmente decida si aplicar o no esta técnica: unos resultados en general mejores pero a costa de obligar al usuario a esperar unos pocos segundos más.

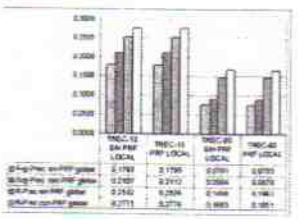

Figura 3: Impacto de la realimentación global

4. *Conclusiones y Trabajo Futuro*

Los sistemas DIR necesitan fusionar listas de documentos de varias colecciones. Normalmente, los algoritmos de fusión de docu-

mentos tienen en cuenta sólo los términos originales de la consulta para re-evaluar cada documento. Por otro lado, PRF es una técnica bien conocida para expandir la consulta original sin interacción con el usuario. Los experimentos que hemos realizado van en esta línea cuando la PRF se aplica a nivel local. No obstante, los resultados que hemos obtenido son muy diferentes cuando aplicamos PRF a nivel global, alcanzando una notable mejora (sobre el 20 %). Para estudiar la aplicación de PRF a nivel global, hemos aplicado el algoritmo 2-step RSV en dos escenarios DIR distintos, uno formado por 13 subcolecciones y otro que consta de 80 subcolecciones. 2-step RSV crea un índice en tiempo de consulta basado el vocabulario de la consulta y los documentos recuperados localmente. En este trabajo se han obtenido mejores resultados aplicando PRF a nivel global, no local.

Un trabajo futuro es el estudio de PRF enviando la consulta expandida a nivel global sobre cada colección. Así la consulta original del usuario es enviada a cada colección seleccionada. Entonces, el 2-step RSV se aplica y la consulta original del usuario es expandida. Esta consulta expandida se envía a cada colección. De esta forma esperamos mejorar los resultados locales. Finalmente, también estamos interesados en las capacidades del algoritmo 2-step RSV sobre otros escenarios DIR: con más colecciones, con mayor número de modelos de pesado distintos y con diferentes tamaños.

5. Agradecimientos

Esta investigación ha sido financiada por el Ministerio de Ciencia y Tecnología (MCYT) con el proyecto TIC2003-07158-C04-04.

Bibliografía

Callan, J. P., M. Connell, y A Du. 1999. Automatic discovery of language models for text databases. En *ACM-SIGMOD International Conference on Management of Data*, páginas 470–490.

Callan, J. P., Z. Lu, y W. B. Croft. 1995. Searching distributed collections with inference networks. En *Proceedings of the 18th International Conference of the ACM SIGIR'95*, páginas 21–28, New York. The ACM Press.

J. J. Rocchio, Jr., 1971. *The Smart Retrieval System: Experiments in Automatic Document Processing*, capítulo Relevance feedback in information retrieval, páginas 313–323. Prentice Hall.

Liu, K., C. Yu, W. Meng, A. Santos, y C. Zhang. 2001. Discovering the representative of a search engine. En *Proceedings of 10th ACM International Conference on Information and Knowledge Managemenet (CIKM)*.

Martínez-Santiago, F., M. Martín, y L.A. Ureña. 2003. SINAI at CLEF 2002: Experiments with merging strategies. *Advances in Cross-Language Information Retrieval. Lecture Notes in Computer Science. Springer Verlag*, páginas 187–197.

Martínez-Santiago, F., A. Montejo-Ráez, L.A. Ureña, y M.C. Diaz. 2003. SINAI at CLEF 2003: Merginig and decompounding. En *Proceedings of the CLEF 2003*, páginas 99–107.

Ogilve, P. y J. Callan. 2001. The effectiveness of query expansion for distributed information retrieval. En *Proceedings of the Tenth International Conference on Information Knowledge Management (CIKM 2001)*, páginas 193–190.

Robertson, S. E, S. Walker., y M. Beaulieu. 2000. Experimentation as a way of life:Okapi at TREC. *Information Processing and Management*, 1(36):95–108.

Robertson, S.E. y K. S. Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146.

Si, L. y J. Callan. 2003a. Distributed information retrieval with skewed database size distributions. En *Proceedings of the National Conference on Digital Government Research*.

Si, L. y J. Callan. 2003b. Relevant document distribution estimation method for resource selection. En The ACM Press., editor, *Proc. of the 26 Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto.

Voorhees, E., N. K. Gupta, y B. Jhonson-Laird. 1995. The collection fusion problem. En *Proceedings of TREC-3*, volumen 500-225, páginas 95–104. National Institute of Standards and Technology, Special Publication.