CAPÍTULO 6

RECUPERACIÓN DE INFORMACIÓN CON *BLINDLIGHT*

El término recuperación de información (IR) hace referencia, en general, al estudio de sistemas automáticos que permitan a un usuario determinar la existencia o inexistencia de documentos (esto es, textos) relativos a una necesidad de información formulada habitualmente como un pequeño fragmento de texto conocido como "consulta" (una o más frases o una simple secuencia de "palabras clave"). Los orígenes de este campo pueden remontarse a los años 1950 y desde entonces ha madurado enormemente, en particular en lo tocante a la evaluación sistemática de los sistemas IR. En la actualidad la recuperación de información está caracterizada por tres atributos dispares: su naturaleza interactiva, el número de documentos a tratar y el carácter multilingüe de documentos y usuarios. En este capítulo se analizará el uso de blindLight como técnica de recuperación de información, las características de la misma que la hacen especialmente interesante en entornos multilingües y los resultados obtenidos en pruebas estandarizadas. Así mismo, se compararán estos resultados con los alcanzados por otras técnicas IR y se señalarán distintas líneas de trabajo futuro que se espera contribuyan a mejorar esta nueva técnica con el objeto de ponerla al mismo nivel que otros métodos de recuperación de información ya afianzados.

1 Recuperación de información

El término "recuperación de información" (*information retrieval* o *IR*) hace referencia al conjunto de procesos necesarios para representar, almacenar, buscar y encontrar información relevante para las consultas de los usuarios (Ingwersen 1992, p. 49). A pesar de su carácter central en el ámbito de las ciencias y tecnologías de la información (Griffith 1980, p. 239) (Järvelin y Vakkari 1992, citado por Ingwersen 1992) se trata de un término vagamente definido (van Rijsbergen 1979, p. 1) puesto que, en principio, podría referirse a diversas manifestaciones de la información como imágenes, audio, texto, etc. No obstante, se acepta generalmente que la "recuperación de información" se ocupa únicamente de información textual (Ingwersen 1992, p. 50) y describe sistemas análogos a los desarrollados,

por ejemplo, por Salton, Spärck-Jones o Robertson. Es interesante en este sentido la definición propuesta por Lancaster (1968, citado por Rijsbergen 1979, p. 1):

> *Recuperación de información es el término que se aplica habitualmente, aunque de manera inexacta, al tipo de trabajo descrito en esta obra*[1]*. Un sistema de recuperación de información no informa al usuario acerca del tema de su consulta, es decir, no modifica sus conocimientos. Simplemente, indica la existencia (o inexistencia) y localización de documentos relativos a dicha consulta.*

Se trata de una definición conveniente puesto que describe a la perfección toda una serie de sistemas desarrollados durante un período de tiempo muy amplio: desde *SMART* (Salton y Lesk 1965) (Salton 1968) hasta *Google* (Brin y Page 1998) por fijar sólo dos hitos. Sin embargo, no tiene en cuenta un aspecto muy importante de la *IR*: su naturaleza interactiva. En fecha muy temprana Don Swanson[2] (1977) señaló:

> *…la recuperación de información es un proceso de ensayo y error… Una consulta no es más que una suposición acerca de los atributos que se espera tenga el documento deseado. En general, se emplea la respuesta del sistema para corregir esa suposición inicial en posteriores intentos.*

No obstante, en los primeros momentos de la investigación en sistemas *IR* no se establece la interactividad como un requisito sino que se trata de una característica que emerge con la evolución[3] de estos sistemas. Los aspectos interactivos e iterativos en los procesos de recuperación de información no comenzaron a estudiarse hasta los años 80 –por ejemplo, Belkin y Vickery (1985), Croft y Thompson (1987), Bates (1989) o Ingwersen (1992)– y fueron totalmente aceptados en los años 90 reconociéndose la necesidad de investigar no sólo los resultados de los sistemas *IR* sino la forma en que son utilizados por los usuarios (Harman 1996).

Así pues es posible distinguir tres enfoques (Ingwersen 1992) en la investigación de sistemas *IR*: el "tradicional", el "orientado al usuario" y el "cognitivo". En el primer caso el objeto de estudio es la representación de documentos y consultas así como las funciones de "emparejamiento" entre ambos tipos de textos. El enfoque "orientado al usuario" se centra en los componentes humanos de un sistema interactivo de recuperación de información. Por último, el enfoque "cognitivo" trata de desarrollar un enfoque integrador de todos los componentes (automatizados y humanos) de un sistema completo.

En este trabajo se ha estudiado la aplicación de *blindLight* a la recuperación de información desde un enfoque "tradicional". Ciertamente, muchas de las aplicaciones de esta técnica (clasificación, categorización o extracción de resúmenes) podrían ser muy útiles en un sistema interactivo y sería muy interesante estudiar su aplicación. Sin embargo, el autor consideró que este análisis no era uno de los puntos a tratar en esta disertación.

A lo largo de los siguientes apartados se repasará brevemente la evolución de los sistemas *IR*, se estudiará la forma en que se realiza su evaluación (desde el enfoque

---

[1] Lancaster, F.W. 1968, *Information Retrieval Systems: Characteristics, Testing and Evaluation.*

[2] No obstante, quizás la contribución fundamental de Swanson haya sido la propuesta de técnicas para generar hipótesis médicas de manera semi-automática a partir de colecciones bibliográficas (Swanson 1986) (Swanson 1991) (Swanson y Smalheiser 1997) (Smalheiser y Swanson 1998). La última implementación de tales técnicas está disponible en `http://arrowsmith.psych.uic.edu/arrowsmith_uic`.

[3] Según Salton y Crouch (1989, p. 3) son las mejoras en el *hardware* y las interfaces gráficas de usuario las que han posibilitado *"sofisticadas interacciones entre los usuarios y los sistemas [de recuperación de información]."*

tradicional), se describirá la forma en que es posible aplicar *blindLight* a esta tarea y se analizarán los resultados obtenidos con distintas colecciones y en la participación en *CLEF'04*[1].

2 Evolución de los sistemas de recuperación de información

Según Herbert Ohlman (1999) las tecnologías de procesamiento del lenguaje tal y como las entendemos en la actualidad comenzaron su andadura a mediados del S. XX. En esta época se desarrollaron técnicas para construir concordancias[2] (véase Fig. 90) como el "indexado por permutación" (*permutation indexing*) (Ohlman 1957) o la técnica análoga *KWIC* (*keyword in context*) de Luhn (1959, citado por Ohlman 1999), se propusieron técnicas para obtener resúmenes automáticos (Luhn 1958) y, por supuesto, se dieron los primeros pasos hacia los actuales sistemas de recuperación de información (Luhn 1957) y (Baxendale 1958).

cenamiento (600 Mb), la rapidez de **recuperación de información**, la posibilidad de realiz**
es implicaciones en el campo de la **recuperación de información** y de la valoración sobre **
rld Wide Web, sistema basado en la **recuperación de información** a partir de técnicas de h**
. Servicio Internet Localización y **recuperación de información** relacionada con las Cienc**
ctualidad, como la extracción y la **recuperación de información**. Además de comunicaciones**
üística de corpus. 3. Extracción y **recuperación de información**. 4. Gramáticas y formalis**
tauración especiales tales como la **recuperación de información** de seguridad, restauració**
as). - Desarrollar los sistemas de **recuperación de información** clínica para usos asisten**

Fig. 90 Concordancias extraídas del *Corpus de Referencia del Español Actual* de la *RAE*.

Luhn (1957, p. 313) planteó lo que podría considerarse el núcleo básico de los sistemas de recuperación de información:

> *Cuanto mayor sea la coincidencia entre los elementos de dos representaciones y entre las distribuciones de los mismos, mayor será la probabilidad de que representen información similar.*

Para implementar esta idea en un sistema viable de recuperación de información Luhn propone, en primer lugar, extraer concordancias para las distintas palabras de una colección a fin de utilizar esos datos para construir (mediante expertos) familias de "nociones". Posteriormente, dichas nociones serían utilizadas para codificar los documentos de la colección, aunque sólo las nociones principales (las más frecuentes o las utilizadas en títulos, encabezados y resúmenes) aparecerían en la representación final de cada documento.

Para llevar a cabo las consultas Luhn sugiere que el usuario proporcione un documento en el que describa de la forma más detallada posible la naturaleza del problema para el que pretende hallar respuesta. Este documento sería codificado de la misma manera que los pertenecientes a la colección y comparado con éstos: a mayor número de nociones en común mayor similitud entre el documento y la consulta. Luhn también apunta la posibilidad de "expandir" automáticamente la consulta original mediante nociones relacionadas.

Luhn no proporcionó ningun resultado empírico, tan sólo afirmó que un experimento llevado a cabo con 1200 informes técnicos produjo "resultados esperanzadores" y, a la vista de los últimos 50 años, sin duda lo eran. Por otro lado, es importante señalar que Luhn introdujo algunos conceptos relevantes en este campo como la

[1] Recuérdese que el *CLEF – Cross Language Evaluation Forum* es un foro internacional para la evaluación de sistemas de recuperación de información que operen sobre idiomas europeos.

[2] Índice de todas las palabras de un libro o del conjunto de la obra de un autor, con todas las citas de los lugares en que se hallan (*RAE* 2001).

utilización de la frecuencia de los términos, del marcado de partes del habla[1] o la expansión de consultas mediante diccionarios de sinónimos.

El trabajo de Baxendale (1958) se centró en la extracción automática de términos clave para su utilización como índices. Baxendale propuso tres métodos para realizar dicha tarea señalando una serie de puntos importantes: la existencia de "palabras vacías", la relación entre "significatividad" y frecuencia de aparición y la dificultad de evaluar la calidad de los términos extraídos automáticamente al compararlos con otros propuestos por expertos.

Los trabajos anteriores plantearon ideas extremadamente interesantes pero fueron Maron y Kuhns (1960) los primeros en proponer un sistema de recuperación que incorporase de una manera efectiva el concepto de "relevancia" de un documento.

Luhn (1958, p. 313) ya había señalado que el grado de coincidencia entre los términos y entre las distribuciones de los mismos en dos documentos sería un indicador de la probabilidad de que ambos documentos traten temas similares. Sin embargo, no describió ninguna técnica para calcular dichas probabilidades automáticamente.

Por otro lado, Baxendale (1958) estaba interesado en reducir cada documento a un número de términos índice pequeños y, aunque no detalló el modo en que se haría la recuperación, cabe aventurar que la técnica empleada sería una búsqueda booleana. Al emplearse muy pocos términos índice por documento los resultados serían reducidos; sin embargo, se plantearían problemas si los usuarios empleasen términos similares aunque distintos de los extraídos del texto de los documentos[2].

Maron y Kuhns señalaron dos aspectos importantes en el campo de recuperación de información: (1) la idea de relevancia como una cantidad numérica que aunque puede carecer de valor como medida cuantitativa sí resulta útil en términos comparativos y (2) el hecho de que una consulta es, por naturaleza, imprecisa, tan sólo una "pista" sobre las necesidades de información del usuario y que el propio sistema *IR* debe "elaborar" dicha pista.

Maron y Kuhns argumentan que conocida la probabilidad de que un término sea aplicado a un documento y sabida la frecuencia de acceso a cada documento es posible, aplicando el teorema de Bayes, calcular la probabilidad de que un documento en particular sea considerado relevante para un término dado y por extensión para una consulta que combina varios términos. Por otro lado, describen una serie de técnicas que permitirían expandir consultas añadiendo nuevos términos relacionados con los empleados por el usuario, así como la forma de ponderarlos. Aproximadamente en la misma época Lauren B. Doyle (1959 y 1965, citado por van Rijsbergen 1979, p. 108) y H.E. Stiles (1961, citado por van Rijsbergen 1979, p. 108) utilizaron de manera similar la co-ocurrencia de términos.

Además de esto, Maron y Kuhns fueron los primeros en proporcionar resultados empíricos y demostrar de manera rigurosa que un sistema de recuperación de información automatizado era factible. Es cierto que no se trataba de búsquedas en "textos completos" y que los términos índice eran extraídos y ponderados manualmente pero la importancia de las ideas planteadas en su trabajo es indudable.

---

[1] *Sería ventajoso identificar mediante símbolos especiales ciertas clases [de palabras] como sustantivos, adjetivos o nombres.* (Luhn 1957, p. 314)

[2] Posteriormente, Lewis, Baxendale y Bennett (1967) investigarían la forma de determinar estadísticamente relaciones de sinonimia/antonimia.

El modelo vectorial de documentos fue utilizado por primera vez en el sistema *SMART* (Salton y Lesk 1965) y ya se ha descrito con detalle en la página 45 y posteriores, baste tan sólo decir que en este modelo los documentos son representados como vectores en un espacio *n*-dimensional donde los términos son empleados como coordenadas a las que se asigna un peso calculado a partir de la frecuencia de uso de cada término en el propio documento y en la colección completa. Para llevar a cabo la recuperación de documentos relevantes para una consulta es necesario representar dicha consulta como un vector y calcular la similitud entre el vector consulta y los vectores documento empleando medidas de asociación como la función del coseno.

Es necesario decir que el modelo vectorial es más bien una familia de técnicas de recuperación de información: por un lado pueden emplearse toda clase de elementos como términos (*n*-gramas, palabras, raíces, lemas, etc.) y, por otro, aplicarse distintos esquemas de ponderación para dichos términos. Así, una de las variantes consideradas más efectivas es la que emplea *tf*idf* junto con la denominada *pivoted document length normalization*[1] (Singhal, Buckley y Mitra 1996) que trata de evitar la tendencia de la medida coseno a favorecer la recuperación de los documentos más cortos de la colección. Por otro lado, el modelo vectorial también admite la expansión automática de consultas sugerida por Maron y Kuhns (1960): Rocchio (1966) introdujo una técnica que permitía ampliar una consulta original empleando para ello la información proporcionada por el usuario al señalar los documentos relevantes dentro del conjunto de resultados[2].

Otro modelo para recuperación de información es el probabilista del que Maron y Kuhns (1960) fueron pioneros. No obstante, serían Karen Spärck-Jones y Stephen Robertson (1976) los que sentarían unas bases realmente sólidas para su utilización efectiva. En la propuesta de Maron y Kuhns era vital conocer la probabilidad con que un término sería utilizado como índice de un documento dado; sin embargo, dicho peso debía ser establecido por un experto humano y el número de términos índice era, por tanto, reducido.

Spärck-Jones (1972) demostró que la especificidad de un término era inversamente proporcional a su frecuencia de uso en la colección, es decir, cuanto mayor es el número de documentos que incluyen un término menos específico resulta como índice y viceversa. Así pues, resultaba posible ponderar cualquier término empleado en una colección y emplearlo como índice de una manera totalmente automática, para un término *t* que apareciese en *n* documentos de una colección formada por *N* documentos el peso sería:

$$w = \log \frac{N}{n}$$

Esta idea sería aplicada, como ya se ha dicho con anterioridad, al método de ponderación conocido como *tf*idf* muy empleado en el modelo vectorial pero, además, daría lugar al mencionado modelo probabilístico que, además de la frecuencia *idf*, utiliza información sobre la relevancia de los documentos para los distintos términos a modo de "entrenamiento" (Robertson y Spärck-Jones 1976) (Spärck-Jones 1979).

---

[1] Normalización de la longitud del documento mediante pivote.

[2] El principal inconveniente de esta técnica es la necesidad de una realimentación explícita que los usuarios son reacios a proporcionar (Balabanovic 1998, p. 6). Una solución a este problema es el denominado *pseudo-relevance feedback* (pseudo-realimentación de relevancia) consistente, *grosso modo*, en la expansión de la consulta original mediante términos extraídos de los primeros documentos obtenidos como resultados. Buckley *et al.* 1994, Robertson *et al.* 1994 o Mitra, Singhal y Buckley 1998 son algunos de los que han mostrado la utilidad de este método.

Este modelo emplea los siguientes parámetros para estimar el peso de un término *t* que aparezca en una consulta *q*: *n* es el número de documentos que incluyen *t*, *N* el número de documentos de la colección, *r* el número de documentos relevantes que incluyen *t* y *R* el número de documentos relevantes para *q*. Robertson y Spärck-Jones proponen distintos esquemas de ponderación aunque el preferido, en particular para situaciones predictivas[1], sería el siguiente:

$$w = \log \frac{(r+0{,}5)(N-n-R+r+0{,}5)}{(R-r+0{,}5)(n-r+0{,}5)}$$

Posteriormente, este modelo sería extendido hasta convertirse en el conocido como *BM25* (Robertson *et al.* 1994) y que es considerado como uno de los métodos *IR* probabilistas más efectivos. Por otro lado es necesario señalar que van Rijsbergen (1977), Harper y van Rijsbergen (1978) o Bookstein y Kraft (1977, citado por van Rijsbergen 1979, p. 108) trabajaron en modelos probabilistas de recuperación de información que no suponen, como el anterior, la independencia entre los términos o que Croft y Harper (1979) desarrollaron un modelo similar pero que no requiere información (explícita o implícita) sobre la relevancia de los documentos.

Naturalmente hay muchos otros modelos para llevar a cabo recuperación de información. Ya se citó el booleano, además del booleano extendido (Salton, Fox y Wu 1983), el vectorial generalizado (Wong, Ziarko y Wong 1985), el basado en conjuntos difusos (Kraft y Buell 1983) o (Cross 1994), o el de semántica latente (Deerwester *et al.* 1990) por citar unos pocos. Para una revisión más profunda de los distintos modelos de *IR* el lector puede acudir al segundo capítulo del excelente *"Modern Information Retrieval"* (Baeza-Yates y Ribeiro-Neto 1999).

3 Evaluación de sistemas de recuperación de información

Un sistema de recuperación de información ideal debería proporcionar tan sólo documentos relevantes para las consultas que recibiese. Sin embargo, en la práctica se acepta que el objetivo de un sistema *IR* es localizar el mayor número posible de documentos relevantes junto con el menor número posible de documentos irrelevantes. Además, el sistema ofrece los resultados de manera ordenada, esto es, cuanto más relevante se presume un documento antes aparecerá en la lista de resultados y viceversa.

Ciertamente, la relevancia de un documento es una cualidad subjetiva que cambia con cada usuario. No obstante, esto no plantea mayores problemas en un marco experimental pues es posible proporcionar una colección de documentos junto con un conjunto de consultas de prueba para las cuales un grupo de expertos decide qué documentos son relevantes. En general, se acepta que si un sistema de recuperación de información funciona de manera adecuada bajo diversas condiciones experimentales también lo hará en condiciones no controladas (van Rijsbergen 1979, p. 113).

Se plantea entonces una serie de cuestiones: la elaboración de colecciones y documentos, las medidas que se tomarán como indicadores del rendimiento, así como el desarrollo del propio experimento. No parece adecuado entrar en excesivos detalles sobre la evaluación en *IR* puesto que el capítulo séptimo de *"Information Retrieval"* (van Rijsbergen 1979) y el resto de referencias de este apartado cubren los distintos aspectos de la evaluación

[1] Obteniendo información sobre la relevancia de los resultados del propio usuario o empleando métodos de pseudo-realimentación de relevancia (que suponen que los primeros *T* resultados son relevantes).

en recuperación de información. Por ello, nos limitaremos a describir someramente las medidas de rendimiento más habituales en *IR* y algunos de los principales esfuerzos hechos para lograr entornos de evaluación válidos.

3.1 ¿Cómo medir el rendimiento de un sistema *IR*?

Dejando a un lado los aspectos interactivos, son dos los parámetros generalmente empleados para evaluar la efectividad de un sistema de recuperación de información: la **exhaustividad** (*recall*) y la **precisión**. La exhaustividad es la proporción de documentos relevantes en la colección que se retornan como respuesta a una consulta mientras que la precisión es la proporción de documentos retornados que son realmente relevantes. Otra medida alternativa es el ***fallout***, la proporción de documentos no relevantes en la colección que aparecen en los resultados (véase Tabla 21).

Estos valores se determinan para cada consulta. Dada una consulta, un sistema *IR* puede retornar una lista de documentos $(d_1, d_2, \ldots, d_k)$ ordenados por relevancia, donde k variaría entre 1 y N, siendo éste el número de documentos en la colección. De este modo se puede determinar la precisión en k y, a partir de dichos valores obtener la denominada **precisión media**. Por otro lado, es posible obtener la llamada **precisión interpolada** para cada consulta en una serie de valores de exhaustividad prefijados (0, 0.1, 0.2, …, 1) y, posteriormente, macropromediar los resultados que se mostrarán en una única **curva precisión-exhaustividad**.

	Recuperados	No recuperados	
Relevantes	w	x	$n_1=w+x$
No relevantes	y	z	$n_2=y+z$
	$n_3=w+y$		$N=w+x+y+z$

Tabla 21. Tabla de "contingencia" en recuperación de información.

En esta tabla *w* es el número de documentos relevantes obtenidos, *x* el número de documentos relevantes que no aparecen en los resultados, *y* el número de documentos irrelevantes que sí aparecen en los resultados y *z* los que no aparecen en los resultados. De este modo la precisión sería w/n_3, la exhaustividad w/n_1 y el fallout y/n_2.

Otra medida que también trata de plasmar la efectividad de un sistema de recuperación de información mediante un valor único es la denominada ***medida F*** de van Rijsbergen (1979, pp. 129-135). En realidad van Rijsbergen propuso un marco en el que podían obtenerse distintas medidas cambiando un parámetro α. La medida original de la efectividad, E, era la siguiente (P es la precisión y R la exhaustividad, *recall*):

$$E = 1 - \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\left(\frac{1}{R}\right)}$$

La medida F es igual a 1-E de tal modo que valores pequeños están asociados a un rendimiento pobre y valores elevados a un rendimiento alto:

$$F = \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\left(\frac{1}{R}\right)}$$

De este modo, si α vale 0 la medida *F* equivaldría a la exhaustividad mientras que si valiese 1 pasaría a ser la precisión. No obstante, el valor habitualmente empleado al referirse a esta medida es α=½ con lo que la medida *F* queda como:

$$F = 2\frac{P \cdot R}{P + R}$$

3.2 Hitos en la evaluación de los sistemas *IR*

Según Harman (1993) el origen de la evaluación experimental en *IR* puede remontarse al trabajo de Cleverdon (1962, citado por Harman 1993) en el proyecto *Cranfield I* donde se evaluaron distintos lenguajes de indexado. Posteriormente, Cleverdon *et al.* (1966) demostraron que las técnicas de indexado automático proporcionaban resultados análogos al indexado manual y sentarían las bases de la evaluación en recuperación de información: la creación de colecciones de documentos, conjuntos de consultas y subconjuntos de documentos relevantes a fin de facilitar la comparación entre distintas técnicas y sistemas.

Spärck-Jones y van Rijsbergen (1975) señalaron que el principal problema de las pruebas elaboradas hasta aquel momento era su limitado tamaño (pocos documentos y consultas) y apuntaron la necesidad de nuevas[1] y mayores colecciones así como el modo de generar las listas de documentos relevantes para la posterior evaluación (el conocido método de *pooling*). El trabajo editado por Spärck-Jones (1981) constituye un punto de inflexión al revisar el trabajo realizado hasta finales de los años 1970 y esbozar las líneas a seguir en las décadas posteriores.

A lo largo de los años siguientes comenzaron a desarrollarse conferencias que pretendían, por un lado, ofrecer un marco de experimentación común a fin de poder comparar distintos sistemas y, por otro, crear colecciones de un tamaño similar a las que debería afrontar un sistema real. Así, en 1992 tuvo lugar la primera edición de *TREC –Text REtrieval Conference*, en 1999 la primera de *NTCIR – NII-NACSIS Test Collection for IR Systems* y en 2000 la primera edición de *CLEF – Cross Language Evaluation Forum*.

TREC proporcionó en su primera convocatoria colecciones de entrenamiento y prueba que contenían alrededor de 1GB de texto en lengua inglesa y consistió en dos tareas: consulta y filtrado de información (Harman 1993). En sucesivas ediciones se incluirían nuevas tareas e idiomas, incluyéndose búsquedas bi y multilingües[2].

NTCIR comenzó de manera similar a *TREC* aunque dirigida al idioma japonés y con una colección de 330.000 documentos (Kando 1999). Al igual que *TREC* con el tiempo se incluyeron otros idiomas de interés en el ámbito nipón (p.ej. coreano, chino e inglés) además de tareas de recuperación de información bi y multilingüe y tareas como la extracción de resúmenes automáticos.

---

[1] Poco después Edward Fox (1983) describiría dos nuevas colecciones, *CACM* e *ISI* (o *CISI*), que, sin embargo, continúan la tradición de colecciones "pequeñas" puesto que la mayor contiene sólo 3.204 documentos. Spärck-Jones y Webster (1979) construirían una de las primera colecciones razonablemente grandes, la *NPL*, con casi 11.500 documentos.

[2] Un sistema *IR* multilingüe permite consultar una colección que contiene documentos escritos en distintos idiomas utilizando cualquiera de los mismos en las consultas y obteniendo resultados que incluirán, naturalmente, varios idiomas. Un sistema bilingüe no es más que una simplificación de este caso general.

CLEF, aunque incluye tareas de búsqueda monolingüe, nació teniendo como objetivo la recuperación de información en entornos multilingües debido al contexto europeo. En su primera edición (Peters 2001) se dispuso de colecciones de artículos periodísticos de un año completo en inglés, francés, alemán e italiano y temas de búsqueda en 8 idiomas europeos. Posteriormente, y de modo análogo a los otros dos foros de evaluación, se han ido añadiendo nuevas tareas (p.ej. respuesta de preguntas, búsqueda de imágenes, de documentos sonoros o en la Web) y se ha ampliado el número de idiomas tanto en las colecciones como en los temas de consulta.

En la actualidad las colecciones de prueba existentes para evaluar sistemas *IR* cuentan con cientos de miles de documentos en múltiples idiomas y algunas de las áreas de investigación más activas incorporan aspectos como el multilingüismo, la interactividad en los procesos de búsqueda, la utilización de documentos hipertextuales o la respuesta de preguntas.

4 Utilización de *blindLight* como técnica de recuperación de información[1]

Como se recordará, el autor afirmó como parte de su tesis que la nueva técnica que proponía podía ser empleada como método de recuperación de información. A continuación se describirá el modo en que es posible adaptar *blindLight* a este fin y, posteriormente, se presentarán los resultados obtenidos con la misma al aplicarse sobre colecciones "clásicas" y en un foro de evaluación internacional como es el *CLEF*.

La utilización de *blindLight* como técnica *IR* es muy sencilla, puesto que tanto documentos como consultas son transformados en los correspondientes vectores de *n*-gramas sin ningún tipo de procesamiento previo; en particular, no se realiza *stemming* ni se eliminan las "palabras vacías"[2]. Esto no sólo facilita la aplicación de la técnica a múltiples idiomas sino que, además, las "palabras vacías" contribuyen de manera sustancial al significado de un texto y proporcionan importantes pistas sobre el dominio de conocimiento (Riloff 1995) por lo que no parece adecuado eliminarlas[3].

Por tanto, al igual que en anteriores aplicaciones de esta técnica, lo único necesario es una medida de similitud entre consultas y documentos que estará construida sobre las dos medidas previamente mencionadas Π y P. Como se recordará, P es la relación entre la significatividad total del vector intersección consulta-documento y la del vector documento mientras que Π es la relación entre la significatividad del mismo vector intersección y la del vector consulta. Puesto que el número de *n*-gramas en consultas y documentos son en general muy distintos, los valores de Π y P no son directamente comparables ya que el primero suele ser mucho mayor que el segundo. Por esa razón se han experimentado hasta el momento diversas formas de "normalización" en las distintas medidas de similitud probadas (véase Fig. 91).

---

[1] Este apartado constituye una evolución de las ideas presentadas en Gayo Avello *et al.* (2004b y 2004c).

[2] Lo cierto es que en ninguna de las aplicaciones de *blindLight* se eliminan las palabras vacías por las mismas razones aquí argumentadas.

[3] En las pruebas realizadas se ha comprobado que la eliminación de palabras vacías mejora sustancialmente el rendimiento cuando no se utiliza ningún método que pondere los pesos de los *n*-gramas en función de su distribución en la colección (véase el apartado "Ponderación inter e intradocumental de los *n*-gramas" en la página 145). No obstante, el rendimiento obtenido empleando únicamente dicha ponderación resulta superior al que se alcanza al eliminar palabras vacías y, además, al combinar este tipo de ponderación con la eliminación de palabras vacías la mejora del rendimiento es inapreciable.

$$S_1 = \Pi$$

$$S_2 = \frac{\Pi + norm(\Pi \cdot \mathrm{P})}{2}$$

$$S_3 = \frac{\Pi + \frac{numgrams(doc)}{numgrams(query)}\mathrm{P}}{2}$$

$$S_4 = \frac{numgrams(query \Omega doc)}{numgrams(doc)}\Pi + \frac{numgrams(query \Omega doc)}{numgrams(query)}\mathrm{P}$$

Fig. 91 Medidas de similitud para un sistema *IR* basado en *blindLight*.

query y *doc* son vectores de *n*-gramas que representan una consulta y un documento, respectivamente. *query*Ω*doc* es el vector intersección de ambos vectores. La función *norm* escala los valores que recibe en el intervalo recorrido por Π a fin de hacerlos comparables mientras que la función *numgrams* retorna el número de *n*-gramas (componentes) de un vector.

Una posibilidad muy interesante a desarrollar en el futuro es la utilización de **programación genética** para encontrar medidas de similitud adaptadas a distintos contextos. Después de todo, el número de *n*-gramas en los vectores de consultas, documentos e intersecciones así como los valores Π y P para cada par `(consulta, documento)` son constantes por lo que sería factible obtener dichos datos para una colección estandarizada y emplear los datos de relevancia a modo de "entrenamiento".

La utilización de programación genética para descubrir funciones de ordenación en sistemas *IR* ya ha sido propuesta por Fan, Gordon y Pathak (2004a y 2004b) y permite obtener funciones que mejoran de manera sustancial métodos como *BM25* (Wang *et al.* 2004). No obstante, el autor consideró que esta línea de investigación quedaba fuera del ámbito de este trabajo y este capítulo recogerá tan sólo los resultados obtenidos con las medidas de similitud expuestas arriba.

Así pues, el funcionamiento de un sistema *IR* basado en *blindLight* es, conceptualmente, muy simple:

- Para cada documento de la colección se calcula y almacena un vector de *n*-gramas que lo representa.
- Cuando el sistema recibe una consulta también la representa mediante un vector de *n*-gramas que será comparado con cada vector documento calculando los valores Π y P correspondientes.
- A partir de estos valores es posible calcular una de las anteriores medidas de similitud que se utilizará para ordenar la lista de documentos que satisfacen la consulta. Aquellos documentos más semejantes a la consulta aparecerán antes que otros con una menor similitud.

Ciertamente, una implementación directa de este modo de funcionamiento es muy ineficiente; sin embargo, es posible desarrollar un sistema basado en *blindLight* que emplee técnicas más eficaces (p.ej. un índice invertido de ficheros implementado sobre tablas *hash* en disco).

4.1 *blindLight* como método *CLIR (Cross Language IR)*

Por otro lado, resulta muy sencillo implementar sistemas de recuperación de información multilingües empleando *blindLight* mediante la técnica de "pseudo-traducción" (Gayo Avello *et al.* 2004c). Se emplea el término "pseudo-traducción" puesto que la técnica

no trata de obtener una traducción de las consultas sino vectores que contengan *n*-gramas que aparecerían en unas hipotéticas traducciones.

Para ello se emplea un *corpus* paralelo[1] de los idiomas fuente (*F*) y objeto (*O*) alineado a nivel de sentencias y se procede del modo siguiente (véase Fig. 92). (1) Dada una consulta escrita en el lenguaje fuente, Q_F, se divide en secuencias de palabras de longitud variable (desde una palabra a la consulta completa). (2) Se explora el *corpus F* en busca de sentencias que contengan alguna de dichas secuencias. (3) Cada sentencia (hasta un máximo *k*) encontrada en *F* es reemplazada por su homóloga en el *corpus O*. (4) Para cada una de estas sentencias homólogas en *O* se obtiene un vector de *n*-gramas y todos estos vectores se intersecan empleando el operador **Ω** (descrito en la página 64). (5) Los distintos vectores obtenidos por intersección se mezclan obteniéndose un vector consulta pseudo-traducido.

Puesto que todas las sentencias homólogas encontradas en *O* contienen presumiblemente la traducción del mismo conjunto de palabras del idioma *F* parece razonable suponer que la intersección **Ω** de los correspondientes vectores contendrá un conjunto de *n*-gramas "traducidos". Por otro lado, aquellas palabras de la consulta que no aparecen en el *corpus* fuente son incluidas directamente en la pseudo-traducción. En teoría, este proceso daría lugar a vectores consulta similares a los que se obtendrían a partir de traducciones reales de las consultas originales.

Este método de pseudo-traducción fue puesto en práctica durante la participación del autor en *CLEF 2004* (Gayo Avello *et al.* 2004c). Puesto que en dicha campaña todos los idiomas objeto de estudio (a excepción del ruso) eran lenguas de la Unión Europea se utilizó como *corpus* paralelo el denominado *Europarl*[2] (Koehn) obteniéndose resultados muy interesantes. Así, para comprobar la "calidad" de las pseudo-traducciones se compararon los vectores obtenidos al pseudo-traducir las consultas *CLEF* de castellano a inglés con los vectores correspondientes a las consultas originalmente escritas en inglés resultando que, en promedio, el 38,59% de los *n*-gramas de las pseudo-traducciones aparecen en las traducciones reales y el 28,31% de los *n*-gramas de estas últimas se encuentran en las primeras. No parece necesario decir que este rendimiento debe mejorarse.

Es preciso señalar que esta técnica tiene cierta relación con las propuestas por Pirkola *et al.* (2002) para encontrar palabras equivalentes en distintos idiomas que difieren en su grafía[3] y por McNamee y Mayfield (2003) para "traducir" *n*-gramas (véase Fig. 93). La diferencia entre tales técnicas y la propuesta por el autor es que éste no pretende obtener traducciones ni para palabras ni para *n*-gramas individuales sino generar a partir de un vector de *n*-gramas correspondiente a un texto en un idioma fuente otro vector que contenga aquellos *n*-gramas que con mayor probabilidad formarían parte de una traducción real a un idioma objeto, vector que podría ser enviado directamente como consulta a un sistema *IR* basado en *blindLight*.

Por otro lado, aunque existen similitudes, la técnica de pseudo-traducción aquí presentada es mucho más sencilla que la de traducción de *n*-gramas individuales de McNamee y Mayfield y, al tiempo, ofrece las mismas ventajas que ésta frente a métodos de

---

[1] Un *corpus* paralelo es una colección de textos traducidos a varios idiomas además del original (*EAGLES* 1996).

[2] *European Parliament Proceedings Parallel Corpus 1996-2003*.

[3] Por ejemplo, `Rwanda` y `Ruanda`, `Chechnya` y `Tsetshenia` o `pharmacology` y `farmakologian` (Pirkola *et al.* 2002).

traducción basados en diccionarios. El funcionamiento de ambas técnicas puede compararse en Fig. 92 y Fig. 93.

(1) Consulta original en el idioma F (castellano):

```
Encontrar documentos en los que se habla de las discusiones sobre la reforma de
las instituciones financieras y, en particular, del Banco Mundial y del FMI
durante la cumbre de los G7 que se celebró en Halifax en 1995.
```

(2) Fragmentos de la consulta a buscar en el corpus F:

```
Encontrar
Encontrar documentos
Encontrar documentos en
…
instituciones
instituciones financieras
…
```

(3) Sentencias del corpus F que contienen el fragmento anterior:

(1315) …mantiene excelentes relaciones con las instituciones financieras internacionales…
(5865) …el fortalecimiento de las instituciones financieras internacionales…
(6145) …La Comisión deberá estudiar un mecanismo transparente para que las instituciones financieras europeas…

(4) Sentencias homólogas en el corpus O (inglés):

(1315) …has excellent relationships with the international financial institutions…
(5865) …strengthening international financial institutions…
(6145) …The Commission will have to look at a transparent mechanism so that the European financial institutions…

(5) Intersección Ω de las sentencias de O homólogas de `instituciones financieras`:

```
{' fi', ' in', 'al ', 'anc', 'cia', 'fin', 'ial', 'ina', 'ins', 'ion', 'itu',
 'l i', 'nan', 'nci', 'nst', 'ons', 'sti', 'the', 'tio', 'tit', 'tut', 'uti'}
```

(6) Vector final correspondiente a la pseudo-traducción de la consulta:

```
{'Bank', 'Worl', 'ld B', ..., 'Hali', 'ifax', ..., 'cumb', 'mbre', ..., 'ssio',
'ussi', 'ards', 'ax e', ' IMF', ..., 'FMI ', 'bre ', 'n Ha', ..., ' G7 ', ' by ',
'the ', ..., 't th', 'n th'}
```

Solapamiento entre la hipotética traducción y el vector pseudo-traducido:

```
Find documents about discussions on the reform of financial institutions, and in
particular the World Bank and the IMF, at the G7 summit that took place in
Halifax in 1995
```

Términos del vector pseudo-traducido que no se corresponden con la traducción hipotética:

```
{' by ', ' cum', ' en ', ' FMI', ' la ', ' la r', ' los', ' oth', ' pos', ' pro',
' rar ', ' to ', ' Uni', 'a re', 'annu', 'ards', 'atin', 'atio', 'ax e', 'bili',
'bre ', 'cont', 'cumb', 'e su', 'e to', 'en H', 'en p', 'Enco', 'ere ', 'FMI ',
'her ', 'ibil', 'ilit', 'in t', 'ing ', 'ion ', 'los ', 'mbre', 'ncon', 'nnua',
'nt t', 'ntra', 'nual', 'ontr', 'orma', 'ossi', 'othe', 'ould', 'poss', 'rds ',
'ring', 'rma ', 's of', 'sibi', 'ssib', 'ted ', 'ther', 'trar', 'tten', 'ual ',
'uld ', 'umbr', 'x en'}
```

Fig. 92 Proceso de pseudo-traducción de una consulta.

Los pasos 1 al 6 muestran el modo en que se lleva a cabo el proceso de pseudo-traducción de una consulta de un idioma F (castellano) a un idioma O (inglés). (4) y (5) muestran los *n*-gramas comunes a las sentencias homólogas. Se incluye además la traducción real de la consulta (que no se emplea en el proceso de pseudo-traducción) junto con el solapamiento entre la misma y la pseudo-traducción. Por último se presentan aquellos *n*-gramas incluidos en el vector incorrectamente pues no pertenecen a la traducción real.

Por último, esta técnica de pseudo-traducción puede resultar un campo de trabajo prometedor puesto que, por un lado, no persigue obtener traducciones de textos destinadas a "consumo humano" lo cual la simplifica enormemente y, por otro, existen toda una serie

de aspectos en los que es posible introducir mejoras. Por ejemplo, la intersección de los vectores de sentencias homólogas puede resultar excesivamente simplista, además, deberían analizarse diferentes *corpora* paralelos[1] así como la posibilidad de emplear *corpora* comparables[2].

```
communist party

_comm  commu  ommun  mmuni  munis  unist  nist_  ist_p  st_pa  t_par  _part  party  arty_
mmuna  munau  munau  munau  munis  unist  unist  ist_p  l_re_  rtie_  _part  rtie_  rtie_

parti communiste
```

Fig. 93 Método de traducción de *n*-gramas de McNamee y Mayfield (2003).

McNamee y Mayfield utilizan una técnica que permite asignar a cada *n*-grama de un idioma fuente (en este caso inglés) un único *n*-grama en un idioma objeto (francés). En este ejemplo proporcionado por sus autores se muestra la "traducción" al francés de los *n*-gramas de `communist party` junto y el solapamiento de los *n*-gramas traducidos y una traducción real.

4.2 Ponderación inter e intradocumental de los *n*-gramas

Ya se ha señalado con anterioridad que la relevancia de un término está estrechamente relacionada con el número de documentos que lo contienen: cuanto mayor es este número menos relevante es el término y, viceversa, la relevancia es mayor en el caso de términos poco comunes. Spärck-Jones (1972) fue la primera en presentar una técnica de ponderación, *idf*, basada en esta hipótesis cuya solidez se ha comprobado no sólo desde un punto de vista empírico sino también teórico (Robertson 2004).

Por lo que respecta al modelo *IR* basado en *blindLight* aún no se ha presentado en este trabajo ningún método de ponderación análogo: cada *n*-grama de un documento tiene un peso obtenido exclusivamente a partir del propio documento sin utilizar ninguna información sobre la distribución de los distintos *n*-gramas en la colección. La ventaja de este enfoque radica en que la colección puede crecer indefinidamente sin necesidad de re-indexar puesto que este proceso se realiza una única vez por documento en el momento en que se añade éste a la colección.

No obstante, a pesar de esta ventaja se está desaprovechando información presente en la colección. Por ello resulta importante analizar el modo de obtener tal información así como su influencia en el rendimiento del sistema a fin de valorar la relación coste-beneficio de su aplicación. Así pues, este apartado presentará un método similar a *idf* y proporcionará más detalles acerca de la ponderación de los *n*-gramas dentro de cada documento.

Para cada documento, D_i, de una colección es posible obtener un vector que asocie a cada *n*-grama del texto su significatividad. Es posible entonces obtener un valor promedio[3]

---

[1] Por ejemplo, *OPUS* (`http://logos.uio.no/opus`) o *MULTEXT-East* (`http://nl.ijs.si/ME`).

[2] Un *corpus* comparable es aquel que contiene textos similares en más de un lenguaje o variedad (*EAGLES* 1996). Reinhard Rapp (1999) señala que dicha similitud viene marcada por un dominio de conocimiento común a todos los textos. En general se considera que un *corpus* comparable consta de varias colecciones de documentos de tamaño y temática similares y que proporcionan aproximadamente el mismo número de términos para cada uno de los idiomas implicados. El gran atractivo de los *corpora* comparables es que son relativamente sencillos de construir en comparación con los paralelos. Así, por ejemplo, en el marco del *CLEF* muchos autores han empleado las propias colecciones de noticias como *corpora* comparables (Rogati y Yang 2001), (Cancedda *et al.* 2003) o (Peinado *et al.* 2004).

[3] Estos valores estadísticos son obtenidos para cada *n*-grama dentro de la colección, es decir, no se calcula el valor promedio de la significatividad de un *n*-grama dentro de los documentos que lo contienen sino en todos los documentos aun cuando no lo incluyan.

para la significatividad de cada *n*-grama que aparezca en la colección así como su desviación típica y su coeficiente de variación[1]. Parece razonable suponer que aquellos *n*-gramas más comunes (p.ej. los correspondientes a palabras vacías) no sólo aparecerán en muchos documentos sino que lo harán con significatividades similares por lo que su coeficiente de variación será reducido mientras que los *n*-gramas más raros presentarán un coeficiente de variación mayor[2].

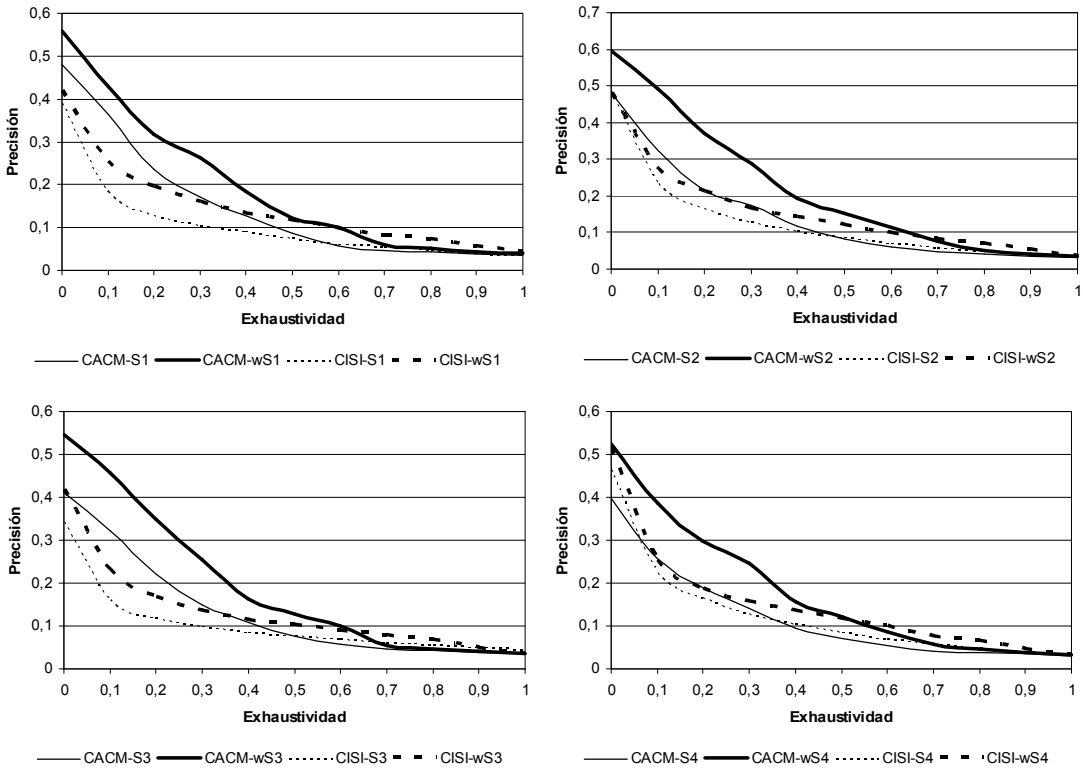


Fig. 94 Influencia de la ponderación interdocumental basada en el coeficiente de variación.

Se presentan los gráficos precisión-exhaustividad para los resultados obtenidos con las colecciones *CACM* y *CISI* empleando las cuatro medidas de similitud presentadas anteriormente. Las medidas no ponderadas están etiquetadas como *sN* (con *N* variando entre 1 y 4) y emplean trazo fino; las medidas ponderadas usan como etiquetas *wSN* y se muestran con trazo grueso. Como era de esperar, en todos los casos la introducción de la ponderación interdocumental supone una mejora sustancial en el rendimiento.

Así pues, al emplear *blindLight* como un sistema *IR*, documentos y consultas se representan mediante vectores de *n*-gramas que tienen asignado como peso el producto de la significatividad del *n*-grama en el documento por su coeficiente de variación en la colección. Como era de suponer, la utilización de información sobre distribución de los *n*-gramas en la colección supone una mejora del rendimiento muy sustancial (véase Fig. 94). Por otro lado, se han estudiado otras formas de ponderación[3], incluyendo *idf*, y en todos los casos el coeficiente de variación ha resultado notablemente superior (véase Fig. 95).

---

[1] El coeficiente de variación es el cociente de la desviación típica entre la media.

[2] El valor máximo que puede alcanzar el coeficiente de variación es $\sqrt{N-1}$ donde N es el tamaño de la colección.

[3] Además de *idf* se ha experimentado con una versión normalizada del coeficiente de variación en el intervalo [0, 1] y con el logaritmo del coeficiente de variación.

Por otro lado, aun cuando a lo largo de este trabajo se ha venido empleando la información mutua para determinar la significatividad, y por tanto los pesos, de los *n*-gramas dentro de cada documento existen otras posibilidades. Por ejemplo, la probabilidad condicional simétrica, los coeficientes Dice y ϕ^2 (Ferreira da Silva y Pereira Lopes 1999) o la ganancia de información (véase Fig. 96).

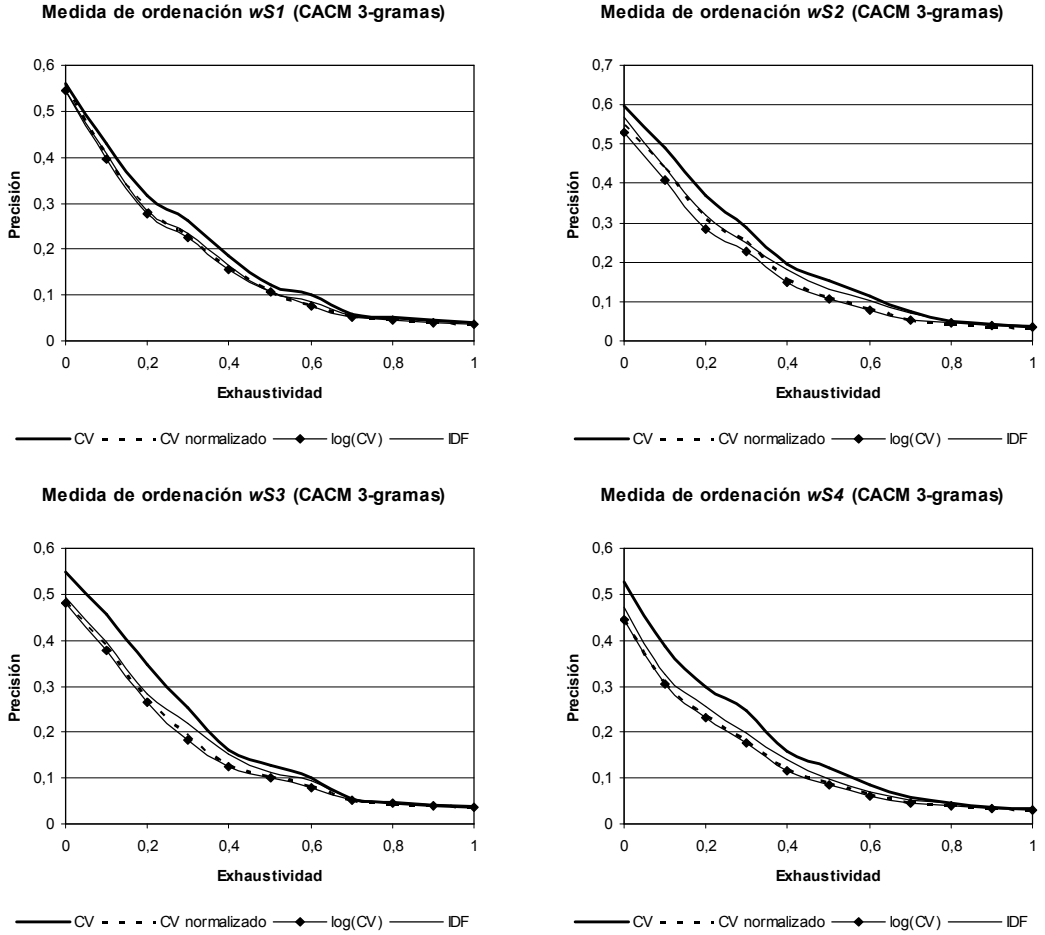


Fig. 95 Distintos métodos de ponderación interdocumental.

Al compararlo con otros métodos (incluido *idf*) el coeficiente de variación resultó el más eficaz siendo las diferencias sustanciales en la práctica totalidad de los casos.

Como se recordará, Ferreira da Silva y Pereira Lopes (1999) desarrollaron una técnica que permitía generalizar una serie de estadísticos para *n*-gramas de longitud arbitraria (véase página 61). Dichos autores utilizaron esas medidas para determinar el grado de "pegajosidad" de *n*-gramas de palabras facilitando así la extracción de términos multipalabra[1]. El autor de esta disertación propuso aplicar la misma técnica a *n*-gramas de caracteres a fin de determinar su grado de significatividad dentro de un texto. En las aplicaciones descritas hasta el momento (clasificación y categorización) se ha empleado la información mutua y en el prototipo participante en *CLEF'04* la probabilidad condicional simétrica.

[1] Simplificando enormemente, cuanto más "pegajoso" resulta un *n*-grama mayor es la probabilidad de que sea un término multi-palabra.

$$Avp = \frac{1}{n-1}\sum_{i=1}^{i=n-1} p(w_1...w_i)\cdot p(w_{i+1}...w_n)$$

$$SI\_f((w_1...w_n)) = \log(\frac{p(w_1...w_n)}{Avp}) \quad (1)$$

$$SCP\_f((w_1...w_n)) = \frac{p(w_1...w_n)^2}{Avp} \quad (2)$$

$$Avx = \frac{1}{n-1}\cdot\sum_{i=1}^{i=n-1} f(w_1...w_i) \qquad Avy = \frac{1}{n-1}\cdot\sum_{i=2}^{i=n} f(w_i...w_n)$$

$$\phi^2\_f((w_1...w_n)) = \frac{[f(w_1...w_n)\cdot N - Avp]^2}{Avp\cdot(N-Avx)\cdot(N-Avy)} \quad (3)$$

$$Dice((w_1...w_n)) = \frac{2\cdot f(w_1...w_n)}{Avx + Avy} \quad (4)$$

$$Infogain((w_1...w_n)) = \frac{1}{n-1}\sum_{i=1}^{i=n-1} p(w_1...w_i)\cdot\log\frac{1}{p(w_1...w_i)} + p(w_{i+1}...w_n)\cdot\log\frac{1}{p(w_{i+1}...w_n)} \quad (5)$$

Fig. 96 Estadísticos para la ponderación de *n*-gramas dentro de un documento.

(w1…wn) es un *n*-grama, *(w1…wi)* y *(wi+1…wn)* son fragmentos consecutivos del mismo (p.ej. para el *n*-grama *'info'* se tendría <*'i'*, *'nfo'*>, <*'in'*, *'fo'*> y <*'inf'*, *'o'*>). *p((w1…wn))* es la probabilidad del *n*-grama *(w1…wn)* en el texto, *p((w1…wi))* es la probabilidad de que un *n*-grama comience con los caracteres *(w1…wi)* y *p(wi+1…wn)* de que termine en *(wi+1…wn)*. *f((w1…wn))*, *f((w1…wi))* y *f((wi+1…wn))* son frecuencias absolutas. *N* es el número de *n*-gramas distintos en el documento. Los estadísticos del (1) al (4) fueron propuestos por Ferreira da Silva y Pereira Lopes (1999) y el quinto por el autor.

A continuación se presentan algunos resultados sobre la influencia del "estadístico de ponderación intradocumental" (véase Fig. 97, Fig. 98, Fig. 99 y Fig. 100); sin embargo, deben considerarse preliminares y es necesario un estudio detallado que combine distintos (1) estadísticos, (2) tamaños de *n*-grama y (3) medidas de similitud (esto es, combinaciones de Π y P). No obstante, puesto que la nueva técnica propuesta no está vinculada a ninguna medida de la significatividad en particular y tan sólo se ha señalado la existencia y viabilidad de varias de tales medidas, un análisis exhaustivo de la eficacia de las mismas queda fuera de los objetivos de este trabajo.

A la vista de semejantes datos no es posible llegar a ninguna conclusión definitiva puesto que, aunque existen diferencias de rendimiento sustanciales, no hay ningun estadístico que resulte claramente superior al resto en todos los casos (es decir, para todas las medidas de similitud). No obstante, la información mutua parece la opción más acertada, especialmente si se combina con el método de ponderación basado en el coeficiente de variación del peso de los *n*-gramas en la colección.

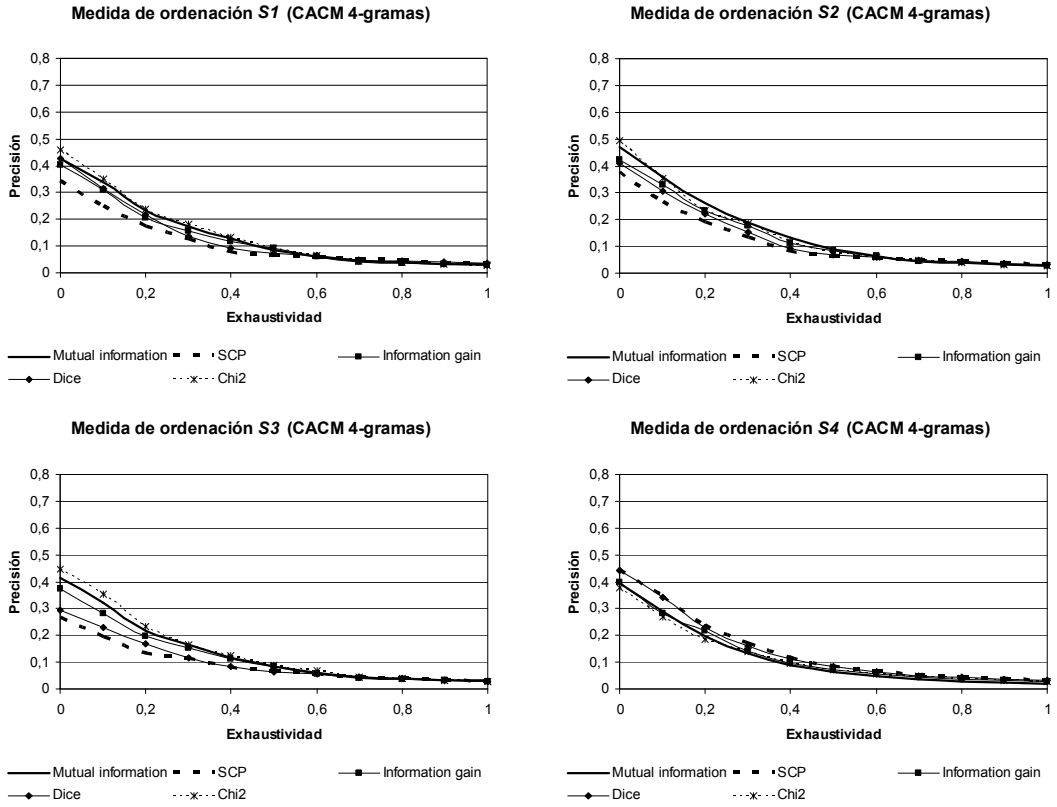


Fig. 97 Rendimiento de los distintos estadísticos para el cálculo del peso de los *n*-gramas.

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	**0,1447**	-3,9%
SCP	0,1142	**-24,1%**
Information gain	0,1352	**-10,2%**
Dice	0,1355	**-10,0%**
ϕ^2	**0,1506**	

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	**0,1553**	
SCP	0,1205	**-22,4%**
Information gain	0,1426	**-8,2%**
Dice	0,1343	**-13,5%**
ϕ^2	**0,1520**	-2,1%

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	0,1389	**-6,1%**
SCP	0,0956	**-35,4%**
Information gain	0,1271	**-14,1%**
Dice	0,1058	**-28,5%**
ϕ^2	**0,1480**	

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	0,1200	**-18,9%**
SCP	**0,1479**	
Information gain	0,1281	**-13,4%**
Dice	**0,1450**	-1,9%
ϕ^2	0,1223	**-17,3%**

Fig. 98 Rendimiento de los distintos estadísticos para la ponderación de *n*-gramas.

(De izquierda a derecha y de arriba abajo, S1, S2, S3 y S4). Como se puede comprobar las diferencias de rendimiento son notables en todos los casos y no puede asegurarse que un estadístico sea claramente superior al resto puesto que los resultados dependen enormemente de la medida de ordenación de resultados utilizada. No obstante, parece que el coeficiente ϕ^2 y la información mutua son los estadísticos que ofrecen de manera consistente los mejores resultados para la mayor parte de funciones de ordenación.

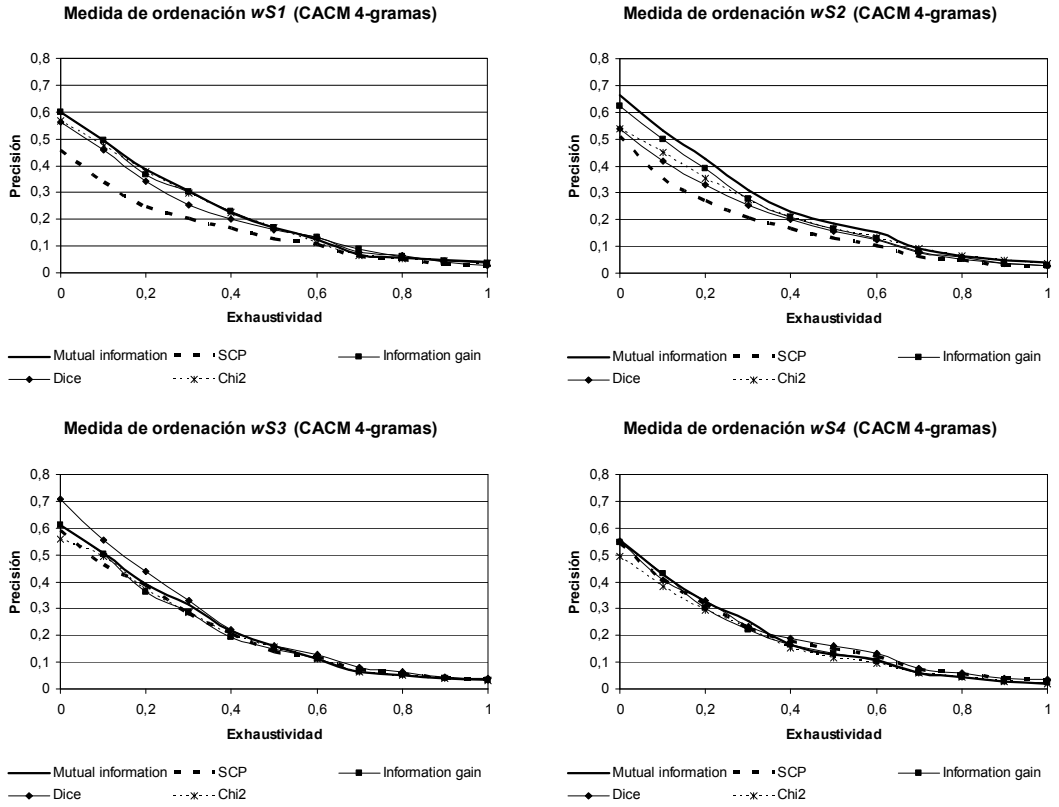


Fig. 99 Rendimiento de los distintos estadísticos para el cálculo del peso de los *n*-gramas (ii).

En este caso se ha utilizado el coeficiente de variación del peso de los *n*-gramas para ponderar los vectores de documentos y consultas.

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	**0,2285**	-0,4%
SCP	0,1636	**-28,7%**
Information gain	**0,2293**	
Dice	0,2108	**-8,1%**
ϕ^2	0,2199	**-4,1%**

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	**0,2492**	
SCP	0,1729	**-30,6%**
Information gain	0,2267	**-9,0%**
Dice	0,2026	**-18,7%**
ϕ^2	0,2150	**-13,7%**

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	0,2275	**-9,7%**
SCP	0,2169	**-14,0%**
Information gain	0,2188	**-13,2%**
Dice	**0,2521**	
ϕ^2	0,2149	**-14,7%**

Ponderación intra-documental	Precisión media 11 pt.	
Mutual information	**0,1919**	-4,6%
SCP	**0,1946**	-3,2%
Information gain	0,1869	**-7,0%**
Dice	**0,2021**	
ϕ^2	0,1744	**-13,3%**

Fig. 100 Rendimiento para los distintos estadísticos para la ponderación de *n*-gramas (ii).

(De izquierda a derecha y de arriba abajo, wS1, wS2, wS3 y wS4). En este caso se ha empleado la técnica de ponderación de *n*-gramas basada en el coeficiente de variación. De nuevo hay diferencias sustanciales en los rendimientos obtenidos. Sin embargo, la información mutua parece perfilarse como uno de los estadísticos más adecuados a la hora de determinar el peso de los *n*-gramas dentro de cada documento.

4.3 Influencia del tamaño de *n*-grama utilizado

El tamaño de *n*-grama utilizado al representar documentos y consultas tiene, como era de esperar, un impacto en el rendimiento aunque menor de lo previsto, dependiendo de la naturaleza de la colección y con resultados "extraños". Así, tanto en la colección *CACM*

como en la colección *CISI* hay un aumento en el rendimiento al pasar de 3-gramas a 4-gramas; sin embargo, mientras que en la colección *CACM* el cambio es sustancial en la colección *CISI* es inapreciable. En cambio, si se comparan los resultados obtenidos en ambos casos al emplear 5-gramas y 3-gramas la mejoría es apreciable (cercana al 10%). Por otro lado, las diferencias entre el uso de 4-gramas y 5-gramas son pequeñas: apreciables en el caso de la colección *CISI* e inapreciables en el caso de *CACM*, aunque, extrañamente, los resultados obtenidos para esta colección empleando 5-gramas son ligeramente peores que utilizando 4-gramas.

En definitiva, el tamaño de *n*-grama tiene una influencia en los resultados obtenidos por el sistema difíciles de evaluar *a priori* en tanto en cuanto parecen venir determinados por la naturaleza de los textos de la colección. Habida cuenta de este hecho, de las distintas posibilidades que se han descrito para la obtención de los pesos de los *n*-gramas en cada documento, así como de las diferentes combinaciones de las medidas Π y P en una única función de ordenación, se abre una línea de trabajo destinada a evaluar de manera sistemática el rendimiento de las distintas configuraciones de *blindLight* al trabajar sobre colecciones e idiomas diversos.

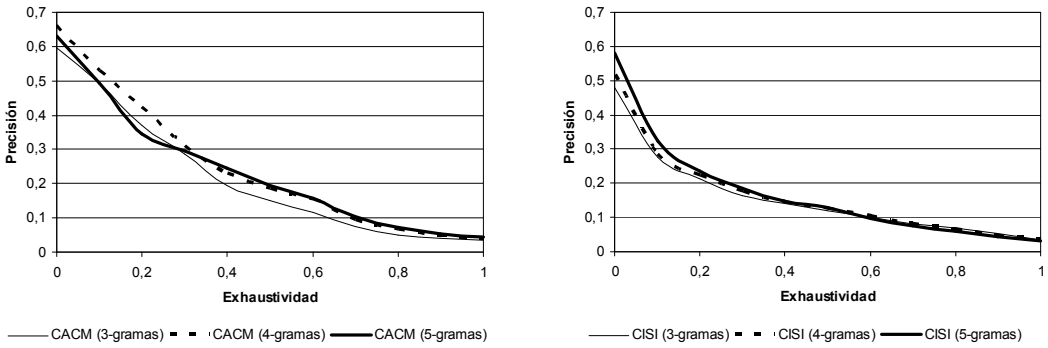


Fig. 101 Curvas *P-R* para las colecciones *CACM* y *CISI* usando distintos tamaños de *n*-grama.

5 Resultados obtenidos por *blindLight*. Comparación con otras técnicas

A fin de estudiar la viabilidad de *blindLight* como técnica de recuperación de información se implementó un prototipo que participó en *CLEF'04* (Gayo Avello *et al.* 2004c). Previamente se realizaron unas pruebas con dos colecciones de menor tamaño, *CACM* y *CISI* (Fox 1983), de las cuales se han presentado algunos resultados a lo largo de los apartados anteriores.

```
.I 47
.W
The use of Bayesian decision models to optimize information
retrieval system performance.  This includes stopping rules to
determine when a user should cease scanning the output of a
retrieval search.
.N 47. Donald Kraft
```

Fig. 102 Una consulta para la colección *CACM*.

La primera colección consiste en un conjunto de títulos y resúmenes de artículos publicados en la revista *Communications of the ACM* entre 1958 y 1979. En total consta de 3204 documentos y 64 consultas (véase Fig. 102 y Fig. 103) junto con los correspondientes "juicios de relevancia". La colección *CISI* consta de 1460 documentos (también resúmenes) y 112 consultas y se proporciona en un formato análogo al de la colección *CACM*. Al preparar los vectores de *n*-gramas para los documentos de ambas colecciones se utilizó el

título, contenido y autor del documento pero no las referencias a otros documentos. Por lo que respecta a las consultas se empleó únicamente el texto de la consulta y nunca el autor de la misma.

```
.I 1457
.T
Data Manipulation and Programming Problems
in Automatic Information Retrieval
.W
Automatic information retrieval programs require the
manipulation of a variety of different data structures,
including linear text, sparse matrices, and tree or list
structures.  The main data manipulations to be performed in
automatic information
systems are first briefly reviewed.  A variety of data
representations which have been used to describe structured
information are then examined, and the characteristics of
various processing languages are outlined in the light of the
procedures requiring implementation.  Advantages of these
programming languages for the retrieval application are
examined, and suggestions are made for the design of programming
facilities to aid in information retrieval.
.B
CACM March, 1966
.A
Salton, G.
.N
CA660315 JB March 3, 1978  11:35 AM
.X
1457       4       1457
1236       5       1457
1457       5       1457
1457       5       1457
1457       5       1457
```

Fig. 103 Un documento de la colección *CACM*.

A la vista de los resultados (véase Tabla 22) es innegable que la actual implementación de *blindLight* como técnica de recuperación de información aún está lejos de proporcionar resultados próximos a los de modelos ya consolidados como el vectorial (Kolda y O'Leary 1998) (Crestani y van Rijsbergen 1998) (Carpineto y Romano 2000) (Tombros *et al.* 2002) (Billhardt *et al.* 2003) o el probabilístico (Crestani y van Rijsbergen 1998). No obstante, su rendimiento es semejante al de técnicas como el indexado mediante semántica latente (Kolda y O'Leary 1998) y puesto que existen diversos aspectos de la propuesta que todavía requieren un análisis exhaustivo es previsible obtener un mejor rendimiento en el futuro.

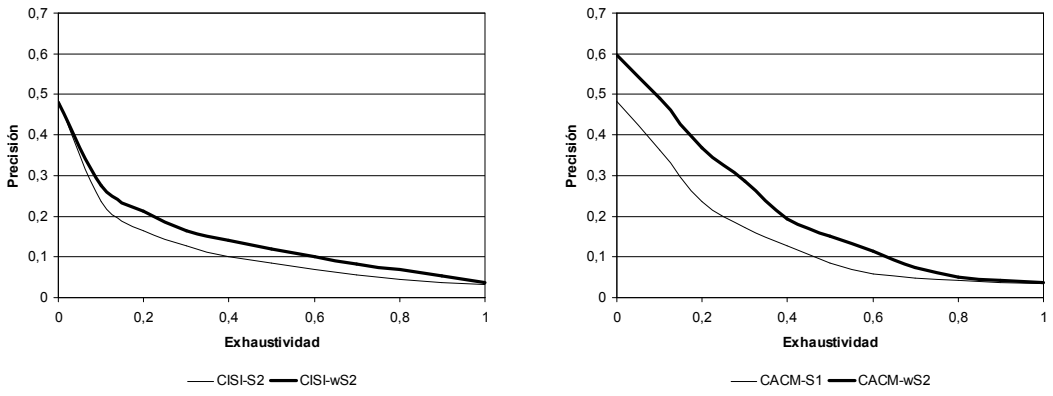


Fig. 104 Resultados obtenidos por *blindLight* sobre las colecciones *CISI* y *CACM*.

	CACM				CISI			
	Precisión media interpolada 11 pt.		Precisión media no interpolada		Precisión media interpolada 11 pt.		Precisión media no interpolada	
blindLight *(wS2, 4-gramas)*	**0,249**		**0,233**		**0,174**		**0,154**	
Modelo vectorial								
(Kolda y O'Leary 1998)	-		-		0,184	5,75%	-	
RbJP (Crestani y van Rijsbergen 1998)	0,271	8,84%	-		-		-	
(Carpineto y Romano 2000)	0,340	36,55%	0,320	37,34%	-		-	
(Tombros *et al.* 2002)	0,378	51,81%	-		0,195	12,07%	-	
(Billhardt *et al.* 2003)	-		0,332	42,49%	-		0,237	53,90%
Hierarchical Clustering (Carpineto y Romano 2000)	0,257	3,21%	0,231	-0,86%	-		-	
Concept Lattice (Carpineto y Romano 2000)	0,281	12,85%	0,253	8,58%	-		-	
Programación genética + CVM[1] (Billhardt *et al.* 2003)	-		0,375	60,94%	-		0,258	67,53%
Modelo probabilístico								
RbLI (Crestani y van Rijsbergen 1998)	0,332	33,33%	-		-		-	
RbCP (Crestani y van Rijsbergen 1998)	0,371	49,00%	-		-		-	
RbGLI (Crestani y van Rijsbergen 1998)	0,428	71,89%	-		-		-	
Semántica Latente								
SDD (Kolda y O'Leary 1998)	-		-		0,181	4,02%	-	
SVD (Kolda y O'Leary 1998)	-		-		0,179	2,87%	-	

Tabla 22. Comparación del rendimiento de *blindLight* en relación con otras técnicas *IR*.

Los resultados obtenidos por *blindLight* aún son sustancialmente inferiores a los alcanzados por modelos como el vectorial o el probabilístico, aunque comparables a los proporcionados por otras técnicas como las de *clustering* jerárquico o semántica latente.

Además, *blindLight* participó en la edición de 2004 del *CLEF* en dos tareas: recuperación de información monolingüe en la colección de documentos escritos en ruso y recuperación bilingüe consultando en castellano la colección de textos escritos en inglés[2]. En el primer caso el prototipo retornó 72 de los 123 documentos relevantes con una precisión media de 0,1433. En la búsqueda bilingüe obtuvo 145 de los 3750 documentos relevantes con una precisión de 0,0644.

5 temas con mejores resultados (ES-EN)		5 temas con mejores resultados (RU)	
218	Andreotti and the Mafia	230	Atlantis-Mir Docking
248	Macedonia Name Dispute	209	Tour de France Winner
202	Nick Leeson's Arrest	210	Nobel Peace Prize Candidates
224	Woman solos Everest	211	Peru-Ecuador Border Conflict
205	Tamil Suicide Attacks	202	Nick Leeson's Arrest
5 temas con peores resultados (ES-EN)		**5 temas con peores resultados (RU)**	
212	Sportswomen and Dopping	227	Altai Ice Maiden
235	Seal-hunting	203	East Timor Guerrillas
241	New political parties	207	Fireworks Injuries
214	Multi-billionaires	228	Prehistorical art
216	Glue-sniffing Youngsters	250	Rabies in Humans

Tabla 23. Temas con los mejores y peores resultados para las tareas monolingüe y bilingüe.

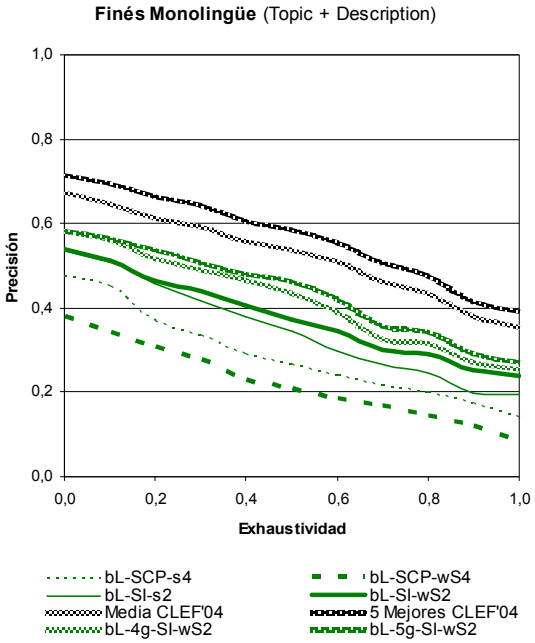
Se han definido los "mejores" como aquellas consultas con la precisión más alta para los 5 primeros documentos retornados y "peores" son las que no retornaron ningún resultado relevante (cuanto mayor era el número de documentos relevantes en la colección peor la consulta). Como se puede ver las consultas relativas a personas, lugares y/o eventos son las que obtienen mejores resultados empleando *blindLight IR* mientras que las consultas abiertas aún no son manejadas de manera adecuada.

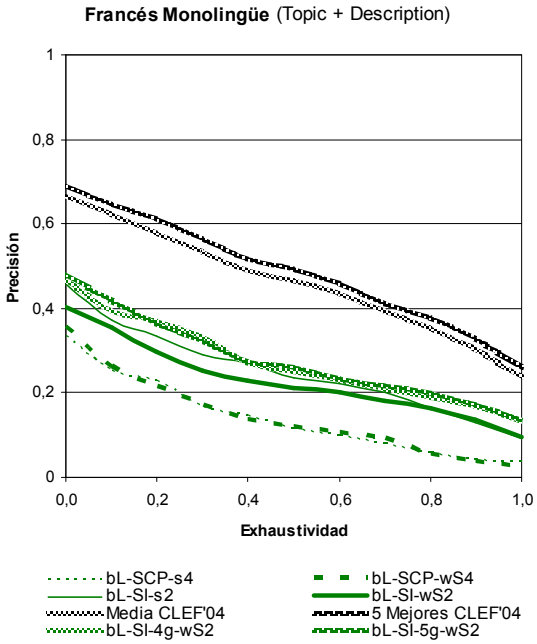
Tales resultados distan mucho de ser buenos pero aun así el autor los consideró alentadores en primer lugar por tratarse de la primera participación en el *CLEF* y en segundo lugar porque aunque el comportamiento promedio es bastante pobre es posible

[1] *Context Vector Model.*

[2] Las consultas escritas originalmente en castellano se pseudo-traducían al inglés y estos vectores de *n*-gramas eran utilizados para consultar la colección de documentos en dicho idioma.

determinar qué clase de temas son los que obtienen peores resultados (véase Tabla 23) señalando una futura línea de trabajo. Hay que señalar, además, que el prototipo participante no empleaba ponderación interdocumental y que el sistema de pseudo-traducción aún está en una fase incipiente todo lo cual influyó sin duda de manera negativa en el rendimiento del sistema en la búsqueda bilingüe.





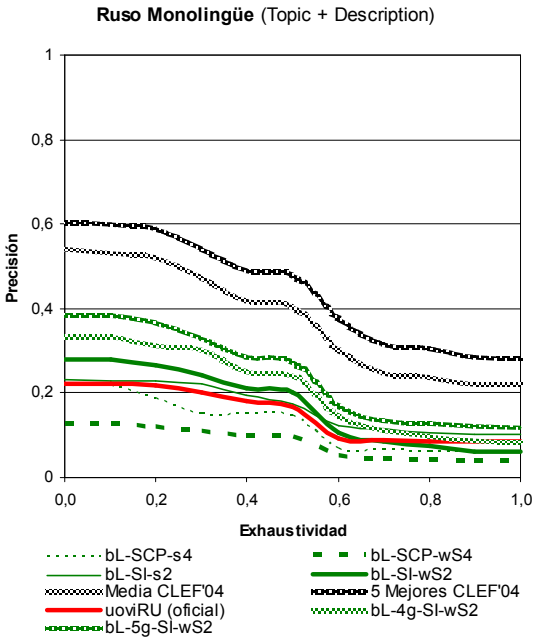


Fig. 105 Pruebas no oficiales de *blindLight* como sistema *IR* monolingüe para finés, francés y ruso.

Las consultas para cada idioma se construyeron automáticamente a partir del título y descripción de los distintos temas (véase Fig. 82). El tamaño de *n*-grama utilizado fue 3 (aunque también se presentan resultados para 4- y 5-gramas) y se experimentó con la probabilidad condicional simétrica y la información mútua como métodos de ponderación intradocumental y con el coeficiente de variación como método de ponderación interdocumental. Para cada combinación de métodos de ponderación se muestra la medida que obtuvo mejores resultados así como los resultados medios de los 5 mejores participantes en *CLEF'04* y los resultados medios de la campaña. En el caso de la prueba en ruso se presenta también el resultado oficial obtenido por el prototipo de *blindLight* (en rojo). Dicho prototipo usaba 3-gramas, *SCP*, la medida de similitud S_2 y no empleaba ponderación interdocumental. Por otro lado, mientras que para el prototipo empleado en *CLEF'04* se transliteró el texto escrito en ruso, el utilizado en estas pruebas puede operar directamente sobre texto *Unicode*. En el caso de finés y francés no hay resultados oficiales pues, aunque inscrito, el autor no obtuvo a tiempo los resultados para su envío a la organización.

Una vez finalizado *CLEF'04* se llevó a cabo una serie de pruebas "no oficiales[1]" de recuperación monolingüe en finés, francés y ruso (véase Fig. 105). En dichas pruebas no sólo se utilizó la probabilidad condicional simétrica (la única empleada en las pruebas oficiales) sino también la información mutua; se experimentó con distintas medidas de similitud además de S_2 y se empleó el método de ponderación interdocumental basado en el coeficiente de variación que se desarrolló con posterioridad a *CLEF'04*.

A la vista de los resultados obtenidos parece que (1) la información mutua resulta sustancialmente mejor que la probabilidad condicional simétrica como método para calcular la significatividad de los *n*-gramas, (2) S_2 resulta sistemáticamente la mejor medida de similitud con independencia del idioma (si se emplea información mutua) y (3) el método de ponderación interdocumental permite, en general, mejorar sustancialmente los resultados. En cuanto al rendimiento de *blindLight* comparado con el de otros sistemas *IR*, estos datos son consistentes con los obtenidos en las pruebas oficiales.

En resumen, la utilización de *blindLight* como técnica de recuperación de información es viable ofreciendo, además, un método de pseudo-traducción de consultas que lo hace muy interesante para entornos multilingües. Ciertamente, los resultados obtenidos por el momento no son tan satisfactorios como los que proporcionan técnicas afianzadas; no obstante, son similares a los de nuevas técnicas consideradas "prometedoras" (p.ej. indexado por semántica latente) y se han señalado una serie de puntos donde, sin duda, será posible mejorar la técnica alcanzando rendimientos superiores.

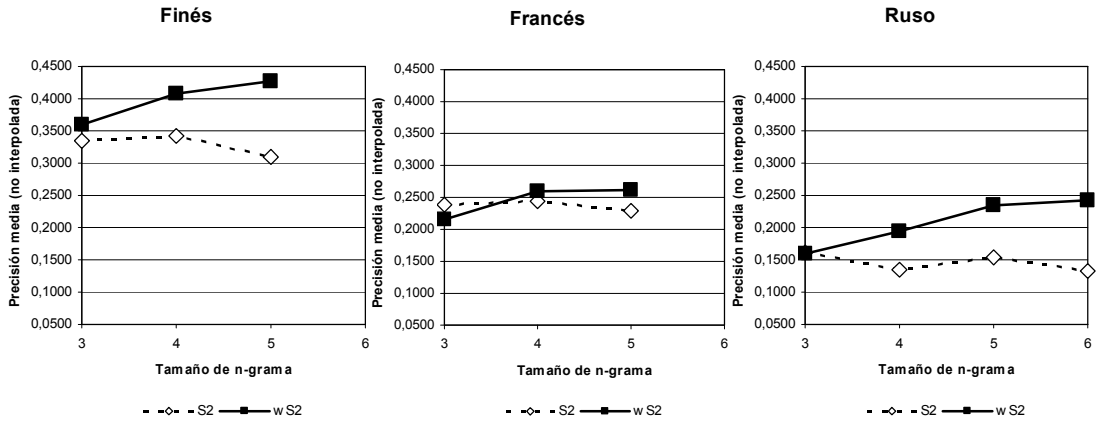


Fig. 106 Evolución de la precisión media (no interpolada) a medida que aumenta el tamaño de *n*-grama empleado para indexar las colecciones.

[1] Son pruebas oficiales aquellas en las que los participantes envían, para ser evaluadas por la organización del *CLEF*, listas de documentos que satisfacen el conjunto de consultas de la campaña en curso. Las pruebas oficiales se llevan a cabo sin que los equipos conozcan los juicios de relevancia. La diferencia entre pruebas oficiales y no oficiales radica tan sólo en el hecho de que en estas últimas es el investigador (y no la organización) quien calcula el rendimiento de su sistema a partir de los juicios de relevancia (mediante el programa `treceval`).