Un modelo de recuperación de información basado en redes bayesianas.

Luis M. de Campos y **Juan F. Huete**
Dpto. de Ciencias de la Computación e I.A. Universidad de Granada. 18071 - Granada.
lci@decsai.ugr.es, jhg@decsai.ugr.es

Juan M. Fernández Luna
Dpto. de Informática. Universidad de Jaén. 23071 - Jaén.
jmfluna@ujaen.es

Resumen En este trabajo presentamos un modelo de recuperación de información probabilístico basado en la construcción de una red bayesiana donde se representan de forma explícita las principales relaciones existentes entre los términos de la colección en uso, dotándolo así de una gran expresividad. La recuperación de documentos se basa en el cálculo de la probabilidad de que cada documento sea relevante dada una consulta y para ello se aplica un proceso de inferencia en dos fases: una propagación exacta en una parte de la red bayesiana más la evaluación de una función en el resto. La capacidad recuperadora del modelo se ha evaluado utilizando varias colecciones estándar de prueba.

1 Introducción.

Una buena definición que establece el proceso completo de *Recuperación de Información (R.I.)* es la que dan Salton y McGill [10], donde indican que la R.I. trata la representación, almacenamiento, organización y acceso de ítem de información. Cuando este proceso se realiza mediante un ordenador, se denomina entonces *Recuperación de Información Automática* y al software correspondiente *Sistema de Recuperación de Información (S.R.I.)*. En nuestro caso, esos ítem de información tomarán la forma de texto, pudiendo ser, por tanto, la representación textual de cualquier objeto.

Dentro de los modelos de R.I. desarrollados destacamos los *modelos probabilísticos* [2] por su clara relación con el modelo que vamos a introducir. Estos estiman la probabilidad de relevancia de un documento a una consulta, siendo la forma en que llevan a cabo esa estimación la diferencia fundamental entre ellos.

En un entorno operacional, es el usuario quien inspecciona los documentos que le ha devuelto el S.R.I. y quien decide la corrección, según el criterio del humano, de la salida, pero en entornos experimentales, se busca una evaluación de la salida mucho más objetiva. Para ello se utilizan dos medidas que nos darán una idea de la calidad de la recuperación: la *exhaustividad*, conocida en inglés como *recall*, que representa la proporción de documentos relevantes recuperados, y la *precisión*, que equivale a la proporción de documentos recuperados que son relevantes. La evaluación de la recuperación se suele hacer calculando la exhaustividad para once puntos estándar de precisión (en caso de utilizar una batería de consultas, la media de esos once puntos para todas las consultas).

Basadas en métodos probabilísticos, las redes bayesianas [8] han sido utilizadas de manera exitosa en entornos caracterizados por la incertidumbre. Este es el caso de la R.I. donde se han aplicado como extensión de los modelos probabilísticos ofreciendo importantes ventajas con las que abordar la incertidumbre intrínseca a los problemas relacionados con esta disciplina [12]. De hecho, la consulta del usuario simplemente es una descripción vaga de su necesidad de información, debido a que, por diversas razones, no es capaz de expresar de una manera precisa lo que realmente está buscando. Otra fuente de incertidumbre es el proceso de construcción de las representaciones de los documentos y consultas ya que da como resultado caracterizaciones incompletas, en forma de listas de términos, del contenido de los objetos representados.

Definamos en primer lugar el concepto de red bayesiana: un grafo dirigido acíclico, donde los nodos representan las variables del problema que se desea resolver. En esta clase de grafos, el conocimiento se representa de dos formas distintas [8]: (a) cualitativamente, mostrando las (in)dependencias existentes entre las variables y (b) cuantitativamente,

expresando la fuerza con que creemos en las relaciones de dependencia, medidas mediante distribuciones de probabilidad condicionadas.

Una vez construida, su uso pasa por la instanciación en la red de un conjunto de variables cuyo valor se conoce y la posterior puesta en marcha de un mecanismo de propagación de probabilidades. El fin último es el cálculo de la probabilidad a posteriori de cada variable del grafo, es decir, la probabilidad de que una variable tome un valor concreto dados los valores que toman las variables instanciadas.

En este trabajo presentamos un modelo de R.I. basado en una red bayesiana. Para ello, son dos los objetivos que vamos a abordar: por un lado, estudiaremos cómo modelar, mediante una red bayesiana, la información documental, tanto desde el punto de vista cualitativo, como cuantitativo. Esta red estará compuesta por dos capas de nodos claramente definidas, representando a los documentos y a los términos que contienen éstos, respectivamente. Por otro lado, estudiaremos el mecanismo que nos permite identificar los documentos de la colección que son relevantes a una consulta de un usuario.

Para poner en práctica estas ideas previas, este trabajo se organiza como sigue: en la siguiente sección comentaremos los principales modelos de recuperación que se basan en las redes bayesianas. Seguidamente, en la sección 3 describiremos detalladamente el modelo de *red bayesiana documental*, centrando la exposición, por un lado en la composición y en la topología del grafo asociado al modelo y, por otro, en los mecanismos que fundan el motor de recuperación. Una vez construida la red en la siguiente sección pasamos a describir cómo se realiza la inferencia con ella. En la sección 5 ofrecemos los resultados de los experimentos que hemos realizado para probar la calidad de nuestro modelo. Para finalizar el último apartado mostrará las conclusiones y propondrá líneas futuras de investigación.

2 *Recuperación de información y redes bayesianas.*

En esta sección vamos a presentar los tres modelos de recuperación de información basados en redes bayesianas más relevantes al modelo que presentamos en este trabajo. El primero, denominado *Inference Network Model*, fue el desarrollado por Croft y Turtle [11]. Está fundado en una red en la que se distinguen a su vez dos subredes: la red de documentos, que es fija para una colección dada, y que contiene básicamente dos tipos de nodos: los términos y los documentos (de los nodos documento salen arcos hacia los nodos término por los que han sido indexados), y la red de la consulta, que se crea cuando el usuario propone una consulta al S.R.I. y contiene nodos consulta y nodos término (los arcos van de los nodos término al nodo consulta). Ambas subredes se conectan por medio de los nodos término que existen en ambas (de los nodos de la red de documentos a la de consultas). Una vez que se han estimado las probabilidades, la inferencia se hace instanciando cada documento sucesivamente y calculando la probabilidad de que la consulta quede satisfecha dado el documento que ha sido observado, es decir, $p(Q \mid d)$. Una vez que todas las propagaciones han finalizado, se genera la correspondiente ordenación de documentos.

Estrechamente relacionado con este trabajo, el de Ghazfan y col. [6] presentan un modelo, básicamente igual que el anterior, pero con la diferencia de que cambian la orientación a los arcos. Formalmente, para una consulta Q los documentos se ordenan según la probabilidad $p(d \mid Q)$. Para ello, se instancian los nodos de la consulta, propagando sólo una vez y calculando así la probabilidad de que cada documento sea relevante dada la consulta.

Por último, Ribeiro y col. [9] presentan el llamando *Belief Network Model*, donde se considera únicamente dos tipos de nodos: documentos y términos, enlazándolos por arcos de los segundos a los primeros. En este modelo, la consulta se considera como un tipo especial de documento. Se propaga también una única vez (como Ghazfan y col.) y se obtiene la ordenación según $p(d \mid Q)$.

Los tres modelos hacen suposiciones de independencia entre términos y, por tanto, no establecen arcos directos entre nodos término. Los modelos Inference Network y Belief Network no aplican ningún algoritmo de propagación como tal, sino que debido a la topología que tienen sus grafos pueden evaluar los valores de probabilidad de manera directa, ofreciendo resultados análogos a los de la propagación.

3 *Descripción del modelo de Red Bayesiana Documental*

El modelo de recuperación que en esta sección presentamos se basa en una red bayesiana a la que vamos a denominar genéricamente *Red Bayesiana Documental*, y cuya descripción se hará en dos partes: comentaremos qué tipos de variables contendrá el grafo subyacente y cómo es su organización interna, es decir, de qué forma se relacionan unas con otras. Una vez especificada la topología, queda la estimación de las distribuciones de probabilidad que almacena cada nodo de la red: los que sean raíces, es decir, los que no tengan padres, albergarán las correspondientes distribuciones de probabilidad marginales y los que sí tengan ancestros, contendrán un conjunto de distribuciones de probabilidad condicionadas. En este momento, la red bayesiana documental está totalmente preparada para ser utilizada en el proceso de inferencia.

3.1 Topología de la Red Bayesiana Documental.

Comencemos, en primer lugar, distinguiendo las variables aleatorias relevantes al problema de la R.I. que estamos tratando, que se corresponden con los nodos del grafo subyacente al modelo : las *variables documento*, que representan a cada uno de los documentos que componen una colección $\mathcal{D}$ y que notaremos como D_j, para $j = 1, \ldots, N$, con N el número total de documentos. Estas variables toman sus valores del conjunto: $\{\bar{d}_j, d_j\}$ (*el documento* D_j *no es relevante* y *el documento* D_j *es relevante*, respectivamente). Por otro lado, *las variables término*, relacionadas con los términos pertenecientes al glosario de la colección, $\mathcal{T}$, las notaremos como T_i, con $i = 1, \ldots, M$, siendo M el número de términos de la colección. Estas variables aleatorias binarias tomarán sus dos posibles valores del conjunto: $\{\bar{t}_i, t_i\}$(*"el término no es relevante"*, *"el término es relevante"*, respectivamente).

El segundo paso es establecer las relaciones que existen entre las variables, es decir, determinar la estructura de la red, para lo cual hacemos las siguientes suposiciones:

1. Para cada término que indexe un documento, existe un enlace entre el nodo que corresponde a ese término y el nodo asociado al documento que indexa.
2. Las relaciones entre documentos solo se dan a través de los términos que contienen dichos documentos.
3. Los documentos son condicionalmente independientes dados los términos por los que han sido indexados. Así, si conocemos los valores de relevancia (o irrelevancia) para todos los términos que aparecen en un documento D_i, entonces nuestra creencia sobre su relevancia no queda afectada por el conocimiento de que otro documento D_j sea relevante o irrelevante.

Teniendo en cuenta estas suposiciones, se establecen implícitamente restricciones en la estructura de la red: por un lado, los enlaces que unen los términos y los documentos en el grafo quedarán dirigidos desde los términos hacia los documentos (al igual que Ghazfan y col. y Ribeiro y col. en sus modelos). Por otro lado, no se establecen relaciones directas entre los nodos documentos. A partir de dichas suposiciones se puede construir una red bayesiana documental con dos capas de nodos: la de documentos y la de términos. En la de documentos los nodos están aislados entre sí, configurando lo que vamos a dar en llamar *subred de documentos*. Con respecto a los términos, asumiremos que éstos no son independientes entre sí. Esta es una de las principales diferencias de nuestro modelo con los tres anteriormente comentados, alcanzando una mayor expresividad, ya que en ellos no se consideran relaciones directas entre términos. La idea es permitir que ante una consulta $Q = (t_1, \ldots, t_k)$ podamos recuperar también aquellos documentos que, sin estar indexados por los términos que aparecen en Q, sí lo están por algún conjunto de términos estrechamente relacionado con ellos. Por ejemplo, ante la consulta "Modelos de recuperación de información probabilísticos", estaríamos interesados en aquellos documentos que hablen de "modelos de recuperación de información basados en redes bayesianas". Por tanto, el objetivo es construir a partir de las colecciones disponibles una red bayesiana que intente recoger las (in)dependencias reales entre términos que existen en la base de datos documental, creando la denominada *subred de términos*.

El determinar estas relaciones basándose en opiniones suministradas por un experto es una tarea prácticamente inviable debido al

elevado número de términos que se utilizan para indexar una colección. Para solucionar este problema, decidimos aplicar un algoritmo de aprendizaje automático que toma como entrada el conjunto de documentos de la colección y genera como salida un poliárbol (un grafo en el que no existe más de un camino dirigido conectando cada par de nodos). Este algoritmo fue diseñado por los autores en [4] como base para llevar a cabo un proceso de expansión de consultas.

La razón fundamental para restringir la estructura de la subred de términos a un poliárbol es la existencia de un conjunto de métodos de inferencia específicos para poliárboles exactos y eficientes, que se ejecutan en un tiempo proporcional al número de nodos existentes [8].

La Figura 1 muestra la topología de la red bayesiana documental que acabamos de describir, donde los arcos con trazo discontinuo han sido aprendidos a partir de la información almacenada en la base de datos documental.

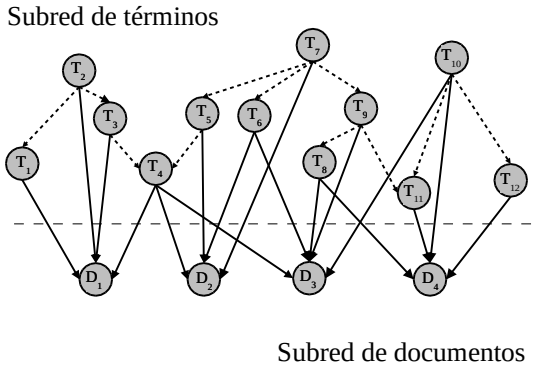

Figura 1: Red bayesiana documental para la recuperación de información.

3.2 Estimación de las distribuciones de probabilidad.

Una vez construida la estructura de la red bayesiana, el siguiente paso antes de poder utilizarla para realizar inferencia es la estimación de las distribuciones de probabilidad que almacena cada uno de los nodos de la red. Así, todos los nodos raíces, es decir, que no tienen padres, deben almacenar las distribuciones marginales. En el caso de la red bayesiana documental, los únicos nodos de este tipo son nodos término, para los cuales hay que estimar $p(t_i)$ y $p(\bar{t}_i)$. Se han diseñado varios estimadores [3], pero el que alcanza un mejor rendimiento es aquel que asigna la misma probabilidad a priori a todos los nodos término raíces, es decir: $p(t_i) = \frac{1}{M}$ y $p(\bar{t}_i) = 1 - p(t_i)$.

Los nodos con padres, tanto términos como documentos, almacenarán el conjunto de distribuciones de probabilidad condicionadas, una para cada una de las posibles configuraciones que pueden tomar los nodos padre. Antes de continuar, establezcamos la notación utilizada: dado un conjunto de términos $T = \{T_1, T_2, \ldots, T_k\}$, definimos una configuración, C, como un vector de la forma $< \mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_k} >$, donde cada uno de los elementos corresponde a un valor que toma cada variable $T_i \in T$. Así, $\mathbf{t_i} = t_i$ si la variable i-ésima de T es relevante y $\mathbf{t_i} = \bar{t}_i$, en caso de que no lo sea. Por ejemplo, para $T = \{T_1, T_2, T_3, T_4\}$, dos posibles configuraciones son $< t_1, t_2, \bar{t}_3, t_4 >$ y $< t_1, \bar{t}_2, t_3, \bar{t}_4 >$. Dado un conjunto de términos T y una configuración C, definimos $n(C)$ como el número de documentos que incluyen todos los términos que figuran como relevantes en la configuración y no incluyen los que figuran como no relevantes en la misma.

Comenzando por los nodos término, éstos albergarán las probabilidades condicionadas $p(T_i \mid \pi(T_i))$, donde $\pi(T_i)$ es una configuración asociada con el conjunto de padres de T_i, notado como $\Pi(T_i)$. Son dos los estimadores que se han diseñado:

- *pc-mv (estimador de máxima verosimilitud)*: utiliza un enfoque frecuentista para calcular los valores de las probabilidades condicionadas, es decir,

$$p(t_i \mid \pi(T_i)) = \frac{n(< t_i, \pi(T_i) >)}{n(\pi(T_i))},$$

  siendo $p(\bar{t}_i \mid \pi(T_i)) = 1 - p(t_i \mid \pi(T_i))$.

- *pc-J (estimador basado en el coeficiente de Jaccard)*: esta fórmula se basa en la medida de similitud de Jaccard [13] la cual, dados dos conjuntos X e Y, calcula la semejanza entre ellos mediante el cociente del número de elementos que componen la intersección y el que forma la unión de ambos conjuntos. Llevada a nuestro contexto, $p(\bar{t}_i \mid \pi(T_i))$ se estima mediante:

$$\frac{n(< \bar{t}_i, \pi(T_i) >)}{n(< \bar{t}_i >) + n(\pi(T_i)) - n(< \bar{t}_i, \pi(T_i) >)}$$

  En este caso, en primer lugar se estima $p(\bar{t}_i \mid \pi(T_i))$ y posteriormente se calcula la probabilidad cuando t_i como $1 - p(\bar{t}_i \mid \pi(T_i))$. Básicamente, la razón

por la que procedemos así se debe a que usando el estimador de Jaccard no se obtiene una distribución de probabilidad, es decir, $p(t_i \mid \pi(T_i)) + p(\bar{t}_i \mid \pi(T_i)) \neq 1$.

Por último, nos queda estimar las probabilidades condicionadas situadas en los nodos documento. Teniendo en cuenta que en una colección de tamaño normal el número de términos que indexan a un documento puede ser de 100 o 200, el número total de combinaciones posibles es enorme. Por tanto, aparecen problemas relacionados con el excesivo tiempo requerido para la estimación, la gran cantidad de almacenamiento necesario y, por último, ya en el proceso de inferencia su utilización por parte del algoritmo de propagación puede causar una ralentización excesiva por cuestiones de acceso a las mismas.

La existencia de estos tres problemas encadenados nos obligó a pensar una forma alternativa a la estimación completa de las matrices, dando como resultado el desarrollo de lo que denominaremos *funciones de probabilidad*. Estas funciones devolverán el valor de probabilidad asociado a una configuración dada en el momento en el que se necesite durante el proceso de inferencia. La función que proponemos en este trabajo se fundamenta en el producto vectorial del vector documento, conteniendo los pesos $tf \cdot idf$ de cada término, con el vector asociado a la configuración $\pi(D_j)$, donde se almacenan los idf de aquellos términos que están a relevantes y 0 cuando no lo están. La función de probabilidad se calcula considerando el producto anterior, una vez normalizado, mediante la siguiente expresión:

$$p(d_j \mid \pi(D_j)) \propto \frac{\sum_{T_i \in D_j, T_i = t_i} tf_{ij} idf_i^2}{k \cdot \sqrt{\sum_{T_i \in D_j} tf_{ji}^2 idf_i^2}}, \quad (1)$$

siendo k la constante de normalización, t_{ij} es la frecuencia del i-ésimo término en el j-ésimo documento y idf_i es la frecuencia documental inversa de dicho término en la colección.

4 El motor de recuperación del modelo.

Una vez que la red bayesiana ha sido construida, ésta se puede usar para obtener un valor de relevancia de cada documento dada una consulta efectuada al S.R.I. por un usuario. En este caso, consideramos que los términos de la consulta actúan de evidencia para el proceso de propagación. Estos términos serán instanciados a "el término es relevante", tras lo cual se ejecutará el proceso de propagación, obteniendo para cada documento d_j la probabilidad $p(d_j \mid Q)$, siendo Q el conjunto de términos que componen la consulta.

Existen dos alternativas básicas a la hora de llevar a cabo la propagación de probabilidades en la red bayesiana [1]: aplicar un mecanismo de inferencia exacto, el cual dependiendo de la topología de la red y del número de nodos que contenga puede hacer que esta fase sea muy costosa en tiempo, o bien, utilizar un algoritmo de propagación aproximado, que obtiene aproximaciones a las probabilidades a posteriori en un tiempo más o menos aceptable. La primera opción, en nuestro caso, y debido a la dimensión del problema es inviable. La segunda ya fue utilizada en [3] con la colección ADI sobre el software de propósito general *Elvira* [5], desarrollado para trabajar con redes bayesianas, pero al extender la experimentación con otras colecciones esta alternativa se mostró inviable en el tiempo. Consecuentemente, y debido a estos problemas se necesita buscar una forma de propagar alternativa, que de forma eficiente nos permita obtener $p(d_i \mid Q)$. La idea básica del método de propagación que proponemos es la siguiente: mediante el algoritmo de propagación exacta en poliárboles de Pearl [8] sólo se propaga en la subred de términos, obteniendo, para cada término T_i de la colección su probabilidad a posteriori $p(t_i \mid Q)$. Con esas probabilidades se evaluará en la subred de documentos la función de probabilidad para obtener para cada documento $p(d_j \mid Q)$.

Esta forma de proceder aplicando la función de probabilidad anterior nos garantiza los mismos resultados que se obtendrían con la propagación exacta en toda la red bayesiana documental [7]. Como resultado de todo el proceso, los grados de relevancia asociados a los distintos documentos se pueden obtener mediante la siguiente expresión:

$$p(d_j \mid Q) \propto \frac{\sum_{T_i \in D_j} tf_{ij} idf_i^2 \cdot p(t_i \mid Q)}{k \cdot \sqrt{\sum_{T_i \in D_j} tf_{ij}^2 idf_i^2}} \quad (2)$$

Nos hemos planteado dos modificaciones a este modelo: una primera consiste es la inclusión de la de la información proporcionada por la frecuencia de los términos de la consulta, qf_i, con objeto de dar más importancia a aquellos términos que se utilizan más frecuentemente. Esta modificación se podría inter-

Col.	Doc.	Term.	Cons.	$\bar{e}_{11p}$	DRR
ADI	82	828	35	0.4706	91
CACM	3204	7562	52	0.3768	246
CISI	1460	4985	76	0.2459	343
CRAN.	1400	3857	225	0.4294	824
MED.	1033	7170	30	0.5446	260

Tabla 1: Características de colecciones y resultados con SMART.

	pc-mv		pc-J	
Col.	Sin qf	Con qf	Sin qf	Con qf
ADI				
DRR	90	91	92	93
$\bar{e}_{11p}$	0.3501	0.4509	0.4130	0.4613
% C.	-25.6	-4.2	-12.2	-2.0
CACM				
DRR	223	247	230	244
$\bar{e}_{11p}$	0.3582	0.4041	0.3759	0.4046
% C.	-4.9	7.2	-0.2	7.3
CISI				
DRR	277	318	282	318
$\bar{e}_{11p}$	0.1948	0.2290	0.2007	0.2301
% C.	-20.8	-6.9	-18.4	-6.4
CRAN.				
DRR	812	814	826	809
$\bar{e}_{11p}$	0.4220	0.4136	0.4314	0.4116
% C.	-1.7	-3.7	0.5	-4.1
MED.				
DRR	288	265	292	266
$\bar{e}_{11p}$	0.6134	0.5836	0.6200	0.5792
% C.	12.6	7.2	13.8	6.4

Tabla 2: Resultados para la expresión (2) sin y con qf.

pretar como una replicación de los términos en la subred de términos a la hora de evaluar la función de probabilidad. Teniendo en cuenta estas restricciones el grado de relevancia de cada documento se obtienen mediante la nueva expresión:

$$p(d_j \mid Q) \propto \frac{\sum_{T_i \in D_j} tf_{ij} idf_i^2 \cdot p(t_i \mid Q) \cdot qf_i}{k \cdot \sqrt{\sum_{T_i \in D_j} tf_{ij}^2 idf_i^2 \cdot qf_i}}$$

Una segunda modificación consiste en ordenar los documentos teniendo en cuenta el incremento de nuestra creencia sobre la relevancia del documento cuando consideramos la consulta como relevante, esto es, disponer los documentos según el valor $p(d_j \mid Q) - p(d_j)$, donde $p(d_j)$ se calcula al propagar en la red sin incluir las evidencias.

5 *Experimentación.*

La fase experimental la hemos puesto en práctica probando nuestro modelo con las colecciones [1] ADI, CACM, CISI, CRANFIELD y MEDLARS, cuyas características principales (número de documentos, términos y consultas) se pueden observar en la tabla 1. En ella también incluimos los resultados que se consiguen con el S.R.I. SMART [10]. Concretamente, mostramos la media de los valores de exhaustividad para los once puntos de precisión ($\bar{e}_{11p}$) y el número de documentos relevantes recuperados (DRR). El esquema de ponderación usado por SMART es *ntc*, es decir, el $tf \cdot idf$ del término normalizado por la raíz cuadrada de la suma de los idf al cuadrado de cada término del vector correspondiente, esquema con el que SMART alcanza un rendimiento bastante alto.

Con qf	ADI	CACM	CISI
DRR	91	242	309
$\bar{e}_{11p}$	0.4581	0.3996	0.2299
% C.	-2.7	6.0	-6.5
Sin qf	CRANFIELD	MEDLARS	
DRR	847	293	
$\bar{e}_{11p}$	0.4421	0.6407	
% C.	2.9	17.6	

Tabla 3: Resultados con *pc-J* y la diferencia de probabilidades del documento.

En una primera fase experimental, el objetivo que pretendemos es doble: por un lado, se busca estudiar el comportamiento del modelo cuando consideramos los mecanismos para estimar las probabilidades condicionadas en los nodos término, es decir, *pc-mv* y *pc-J*; por otro lado, pretendemos estudiar cómo afecta en la capacidad recuperadora la inclusión de los términos de la consulta. Los resultados de esta primera etapa se pueden ver en la Tabla 2, donde incluimos el número de documentos relevantes recuperados en todas las consultas, la media de la exhaustividad para los once puntos de precisión para todas las consultas y el porcentaje de cambio de esa media con respecto al obtenido por el S.R.I. SMART (% C.).

En general, podemos decir que se obtiene un mayor rendimiento en cuatro de las cinco colecciones cuando se hace la estimación de las distribuciones de probabilidad condicionadas de los términos con Jaccard, es decir, *pc-J*. La mejora se hace notable en ADI, CACM y un poco menos significativa en CISI, CRANFIELD y MEDLARS cuando no se

[1] Toda la fase de preprocesado de las colecciones la ha realizado SMART mediante su módulo de indexación.

Modelo	Col.	$\bar{e}_{11p}$	$\bar{e}_{11p} - RBD$
Inf. Net.	ADI	0.2154	0.4709 (113.62)
Inf. Net.	CACM	0.3740	0.4046 (8,18)
Ghazfan	ADI	0.5033	0.4709 (-6.44)
Ghazfan	CACM	0.3730	0.4046 (8.47)

Tabla 4: Comparativa con otros modelos basados en redes bayesianas.

tienen en cuenta los qf e incluso empeora en la mayoría de las colecciones, aunque muy poco, cuando sí intervienen los qf. En esta primera etapa, sólo MEDLARS supera a SMART con *pc-mv* y sin qf. Con qf se une al grupo $CACM$ con un cambio muy significativo. Por otro lado, con *pc-J*, CRANFIELD y MEDLARS son las que sobrepasan a SMART sin usar los qf y se les añade CACM cuando se usan. De manera general, parece que a todas las colecciones salvo a MEDLARS, y finalmente a CRANFIELD, les agrada el uso del qf. Para aquella primera, el decremento en el rendimiento cuando se usa es notorio (casi se pierde la mitad). Deducimos, a la luz de estos resultados, que el uso de la estimación *pc-J* es bastante conveniente. En el caso de *pc-mv*, su uso depende de la colección, al igual ocurre con la incorporación de los qf, ya que en unas es más conveniente que en otras.

Teniendo en cuenta estos resultados, se ha diseñado un segundo grupo de experimentos para analizar el comportamiento del modelo cuando se considera la diferencia de probabilidades a posteriori y a priori de los documentos fijando el estimador *pc-J* para todas las colecciones y la inclusión del qf para ADI, CACM y CISI. Para las dos restantes, no se tiene en cuenta. Los resultados aparecen en la tabla 3.

Se observa como realizando la ordenación de documentos con la diferencia de probabilidades se origina que en CRANFIELD y MEDLARS se mejore sensiblemente. En el resto se mantienen o mejoran los porcentajes de cambio pero relativamente poco.

En cuanto a la comparación de nuestro modelo con los tres restantes basados en redes bayesianas, destacar, según la tabla 4, que en las colecciones ADI y CACM el rendimiento es mayor que el modelo *Inference Network* y con respecto al de Ghazfan y col. se empeora comparando con ADI y se mejora al establecer la comparación con CACM. Los autores del modelo *Belief Network* experimentan con otras colecciones con las que nosotros no hemos probado, y por tanto, no podemos establecer ninguna comparación. Estos resultados comparativos no se pueden tomar de manera absoluta, si no simplemente orientativa, pues las condiciones de experimentación varían en los tres modelos. A pesar de esto, lo que sí se puede concluir es que nuestro modelo presenta un comportamiento que está a la altura de los dos con los que hemos podido comparar.

6 *Conclusiones y trabajos futuros.*

En este trabajo hemos expuesto un modelo de recuperación de información basado en redes bayesianas, cuya principal diferencia con respecto a los ya existentes es la incorporación de las relaciones entre los términos de la colección. La forma de representar esas relaciones ha sido mediante una red bayesiana basada en un poliárbol. Se han planteado dos formas alternativas de calcular las distribuciones de probabilidad condicionada albergadas en los nodos término, así como una solución totalmente válida al problema de la propagación: propagación exacta en la subred de términos y evaluación de una función de probabilidad en la subred de documentos para finalmente calcular $p(d_j \mid Q), \forall D_j$. Por último, hemos presentado dos modificaciones a la misma.

Como primera conclusión de este trabajo, a la luz de los resultados empíricos, podemos decir que el rendimiento del modelo que proponemos depende totalmente de la colección con la que se pruebe, ya que lo que en principio parece que es bueno para unas, no es válido para otras. De manera general, podemos decir que prácticamente todas las colecciones aumentan su eficacia recuperadora cuando se utiliza el estimador de las distribuciones de probabilidad condicionada *pc-J* basado en la medida de Jaccard, que las colecciones ADI, CACM y CISI rinden más al incorporar los qf a la función de relevancia, al contrario que ocurre con CRANFIELD y MEDLARS. El uso de la diferencia de probabilidades de los términos al evaluar la función de probabilidad es algo mejor o aproximadamente igual que su puesta en práctica sólo con la a posteriori del término. En general, podemos concluir que el modelo presentado tiene un mejor comportamiento que SMART y ligeramente superior que los modelos *Inference Network* y el de Ghazfan y col. Un

estudio detallado de las características de las colecciones con que hemos probado nos permitirá comprender y, por tanto, justificar el comportamiento de nuestro modelo de manera más precisa.

Por último, comentar algunas de las líneas futuras de investigación, las cuales se centran en mejorar la calidad de la subred de términos mediante algún proceso de selección de términos. De esta manera sólo se aprendería el poliárbol con los mejores y el resto, los más malos, permanecerían aislados entre sí, evitando así la posibilidad de que introduzcan mucho ruido en la recuperación. Esto aliviará sin duda los procesos de aprendizaje e inferencia, esperando una mejora del rendimiento del modelo. También se pretende incluir las relaciones existentes entre los documentos en la subred de documentos. Y por último, pondremos a prueba nuestro modelo en el proceso de realimentación de relevancia.

Agradecimientos: Este trabajo ha sido financiado por la Comisión Interministerial de Ciencia y Tecnología (CICYT) bajo el proyecto TIC2000-1351.

Referencias

[1] E. Castillo, J.M. Gutiérrez, and A.S. Hadi. *Sistemas expertos y modelos de redes probabilísticas.* Academia de Ingeniería, 1996.

[2] F. Crestani, M. Lalmas, C. J. van Rijsbergen, and L. Campbell. Is this document relevant?... probably. a survey of probabilistic models in information retrieval. *ACM Computing Survey*, 30(4):528–552, 1991.

[3] L. M. de Campos, J. M. Fernández, and J. F. Huete. Building bayesian network-based information retrieval systems. In 11^{th} *International Workshop on Database and Expert Systems Applications:* 2^{nd} *Workshop on Logical and Uncertainty Models for Information Systems (LUMIS)*, pages 543–552. Database and Expert Systems Applications, 4–8 September 2000.

[4] L. M. de Campos, J. M. Fernández, and J. F. Huete. Query expansion in information retrieval systems using a bayesian network-based thesaurus. In *Proceedings of the* 14^{th} *Uncertainty in Artificial Intelligence Conference*, pages 53–60, July 1998.

[5] Dpto. Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada. Elvira. *http://leo.ugr.es/elvira/elvira.html*, 2000.

[6] D. Ghazfan, M. Indrawan, and B. Srinivasan. Toward meaningful bayesian networks for information retrieval systems. In *Proceedings of the IPMU'96 Conference*, pages 841–846, 1996.

[7] Juan Manuel Fernández Luna. *Modelos de recuperación de información basados en redes de creencia.* PhD thesis, E.T.S. Ingeniería Informática. Universidad de Granada., 2001.

[8] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan and Kaufmann, San Mateo, California, 1988.

[9] B. A. Ribeiro-Neto and R. R. Muntz. A belief network model for ir. In H. Frei, D. Harman, P. Schäble, and R. Wilkinson, editors, *Proceedings of the* 19^{th} *Annual International ACM–SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)*, pages 253–260. ACM, 1996.

[10] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc., 1983.

[11] H. R. Turtle and W. B. Croft. Inference networks for document retrieval. In Jean-Luc Vidick, editor, *SIGIR'90,* 13^{th} *International ACM–SIGIR Conference on Research and Development in Information Retrieval, Brussels, Belgium, 5-7 September 1990, Proceedings*, pages 1–24. ACM, 1990.

[12] H. R. Turtle and W. B. Croft. Uncertainty in information retrieval systems. In A. Motro and P. Smets, editors, *Uncertainty management in information systems: from needs to solutions*, pages 189–224. Kluwer Academic Publishers, 1997.

[13] C. J. van Rijsbergen. *Information retrieval. Second Edition.* Butter Worths, London (U.K.), 1979.