UNIVERSIDAD DE GRANADA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

Departamento de Ciencias de la Computación
e Inteligencia Artificial

MODELOS DE RECUPERACIÓN DE INFORMACIÓN BASADOS EN REDES DE CREENCIA

TESIS DOCTORAL

Juan Manuel Fernández Luna

Granada, mayo de 2001

La memoria titulada **Modelos de recuperación de información basados en redes de creencia**, que presenta Juan Manuel Fernández Luna, para optar al grado de DOCTOR, ha sido realizada en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada, bajo la dirección de Luis Miguel de Campos Ibáñez y Juan Francisco Huete Guadix, Profesores Titulares del Departamento en el que se ha realizado la memoria.

Granada, mayo de 2001.

Los directores

Fdo.: Luis Miguel de Campos Ibáñez   Fdo.: Juan Francisco Huete Guadix

El doctorando

Fdo.: Juan Manuel Fernández Luna

A mis padres, Marina y Manolo,
y a mi hermana Marina.

"Un amigo es una persona que nos conoce
muy bien y que, a pesar de eso, nos quiere"
(Helbert Hubbart)

A mi amiga María José.

Agradecimientos

¡Al fin llegaron los agradecimientos! Esto es buena señal, porque, a pesar de que es lo primero que se encontrará el lector de esta memoria, es lo último que escribiré de ella. Y ya tenía ganas, ya... Por un lado, porque representan el colofón a un trabajo de varios años. Por otro lado, debido a que durante todo este tiempo ha sido mucha la gente que me ha apoyado, ayudado y aconsejado y creo que puede ser éste un buen lugar para expresar públicamente mi agradecimiento a todos ellos.

Cuando en el año 94 finalicé mis estudios de Informática tenía claro que quería probar profesionalmente en la Universidad, ya que siempre me ha llamado mucho la atención las dos vertientes que la componen: la docente y la investigadora.

Mis primeros pasos en el mundo laboral los di en la empresa privada, aunque paralelamente en el curso 96-97 logré, tras un primer intento fallido, matricularme en los cursos de doctorado del Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada. Mi tutor era el Dr. Luis Miguel de Campos Ibáñez. A partir de ese momento, los ratos libres que me dejaba mi trabajo los dedicaba a asistir a los cursos, a preparar las tareas que se pedían en ellos y a leer el material que me iba dando mi tutor. Era una afición más y como "sarna con gusto no pica"...

Un día cayó en manos del Dr. de Campos un artículo donde se aplicaban las redes bayesianas a la Recuperación de Información. Le pareció interesante, y visto mi empeño por las cuestiones investigadoras y considerando que podría ser un problema que diera bastante juego, el Dr. Juan Francisco Huete Guadix y él me propusieron adoptarlo como tema para una tesis doctoral. Y así fue cómo a finales de septiembre de 1997 entregamos el proyecto de tesis y nos pusimos manos a la obra con mucha ilusión.

Desde que terminé mis estudios no cejé en el empeño de ser profesor de Universidad. Fue en octubre de 1999 cuando, tras varios intentos de entrar en diferentes universidades andaluzas, lo hice en la de Jaén, donde actualmente soy profesor. En ese momento, la tesis dejaba de ser una afición, configurándose como la condición necesaria para poder continuar así mi carrera universitaria. Pero ese cambio sólo fue conceptual, de denominación, ya que en la práctica, la única consecuencia que realmente ha tenido es que me he podido dedicar mucho más tiempo a algo que realmente me gusta. ¿Se puede pedir más?

Quiero agradecer a los doctores de Campos y Huete la confianza que desde el primer momento han depositado en mí y que les llevó a adoptarme como doctorando. Sólo espero que no les haya decepcionado y que se hayan cumplido la gran mayoría de expectativas que sobre mí pusieron al comienzo y durante todo el desarrollo de este trabajo. Gracias también por la paciencia que han mostrado conmigo (la del Santo Job al lado de la de ellos se queda a la altura de una zapatilla) porque han sido muchos los despistes y las veces que me he equivocado y me he vuelto a equivocar.

Creo que he sido un doctorando afortunado, porque en realidad, más que directores de tesis he tenido compañeros: no todos los directores fomentan el diálogo abierto, la propuesta de ideas y un ambiente de trabajo cordial y relajado; no todos se sientan a tu lado, delante del ordenador, a ayudarte a buscar el error por el cual el programa de turno lleva sin funcionar correctamente una semana; no todos se ponen a echarte una mano (y las dos si hace falta) en la escritura de un trabajo para un congreso; no todos abandonan lo que estaban haciendo en ese momento (incluso asuntos personales) para atenderte cuando llegas con alguna duda o problema; no todos te explican una y otra vez algo que no entiendes hasta que al final lo haces y se tiran contigo el tiempo que haga falta para ello; no todos te están animando continuamente y empujando y tirando de ti en todo momento, (sobre todo cuando está cerca el final y las tareas pendientes se multiplican y todo se ve negro). Y esas, y muchas otras cosas, lo han hecho ellos conmigo. Además, puedo presumir de que estos compañeros son también amigos.

Les quiero felicitar por el trabajo serio y riguroso que han hecho durante todo este tiempo. Ellos se han llevado una gran parte de las muchas horas de estudio y de quebraderos de cabeza que ha tenido esta tesis. Su esfuerzo y dedicación han sido encomiables y dignos de mención. Felicidades por el buen trabajo y GRACIAS POR TODO, Luis, Juan.

Mis padres, Marina y Manolo, y mi hermana Marina también han jugado un papel importantísimo en este período. Siempre he tenido su apoyo incondicional y consejo en cualquiera de las decisiones profesionales que he afrontado. Y durante el tiempo que llevo en la Universidad, donde ha habido temporadas en que me han visto poco y los he tenido muy descuidados, y otras en las que han sufrido especialmente mis nervios y cambios de humor, han sido comprensivos a más no poder. En todo momento me han estado animando para continuar con el trabajo y me han mostrado su interés por conocer su desarrollo. Sin su apoyo, Usted no estaría leyendo esta memoria. Mamá, papá, Marina: gracias.

Mis tíos, Encarnita y Juani, han estado siempre al pié del cañón en este tiempo, interesándose por la evolución del trabajo, aconsejándome en las tesituras en las que me he encontrado, animándome y aportado comentarios interesantes para mejorar el trabajo. Gracias, titos.

El trabajo que se presenta en esta memoria ha sido desarrollado en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada, dentro del seno del grupo de investigación Tratamiento de la incertidumbre en sistemas inteligentes. A los miembros de ambas entidades les agradezco el interés y el apoyo mostrados, así como el uso que he hecho de sus infraestructuras durante este tiempo (mis amigos aliatar, canibal7, santino, moraima, leo, gte, bahía, cóndor -en especial sus discos duros-, hp1, hp3). En este sentido, mención especial merece el Dr. de Campos por sacrificar su intimidad y permitirme que pudiera desarrollar este trabajo en su despacho.

A Serafín Moral Callejón y Andrés Cano Utrera quiero agradecerles su disposición para resolverme cualquier duda y echarme una mano en lo que hiciera falta, el interés que han mostrado por mí, así como su calidad humana y amistad. A Silvia Acid Carrillo tengo que agradecerle una gran cantidad de cosas: desde ser mi amiga, hasta el ánimo, apoyo, consejo, tanto en el campo profesional como en el personal, y cariño que he recibido desde que la conozco, pasando por los paseos por el Albaycín o el software que me prestó para comenzar a funcionar, entre otras muchas cosas. Silvia, muchas gracias.

Gracias también a mis compañeros del Departamento de Informática de la Universidad de Jaén, y muy especialmente a los moradores de la estación orbital "LinaSat".

Sería innumerable citar aquí a todos las amigas y amigos que se han interesado por mi trabajo, me han expresado su apoyo, ánimo y comprensión y con los que he podido contar en los momentos en los que los he necesitado. A todos ellos, muchas gracias. Pero en especial, y como máximos exponentes de ese apoyo y amistad, me gustaría destacar a José María Fernández Garrido, por estar siempre mi lado animándome; a José Manuel Benítez, por ser el mejor administrador de sistemas, por resolverme continuamente problemas, por dejarme sus máquinas, por aconsejarme, por tener una paciencia infinita, por perder su tiempo conmigo, por darme dos tortas cuando me las he merecido, por condonarme (espero) la deuda multimillonaria que he ido contrayendo con él en este tiempo (y no sigo porque se convertiría esto en un monográfico). Gracias, Joses. Y a María José del Jesus Díaz, quien es, simplemente, la çulpable"de que esté próximo a cruzar la meta del maratón en el que más ilusión me hacía participar. Ella me dio el empujón para empezar a correr, me ha estado animando en todo este camino y siempre ha estado ahí para cuando la he necesitado. María José, gracias.

Otro amigo al que tengo que hacer explícito mi agradecimiento es Modesto Jesús Garrido Ruiz, además de por su continuo interés, por su asesoramiento lingüístico para la confección de este documento. Es meritorio que alguien te pueda dar consejos útiles sobre la redacción en español cuando todo parece estar escrito en chino. Gracias, Mode.

Por último agradecer a la CICYT, mediante la financiación de los proyectos TIC96-0781, inicialmente, y TIC2000-1351, actualmente, el apoyo económico que nos ha permitido sufragar la mayor parte de los gastos de este trabajo.

Y como lo breve, si bueno, dos veces bueno, a todos, gracias de corazón.

Índice general

Índice de figuras

Índice de cuadros

Índice de Algoritmos

Notación

a, a_i, b	...	Constantes.
$\mathcal{A}$	...	Conjunto de términos aislados en el poliárbol de la red mixta.
C	...	Configuración.
$\mathcal{C}$	...	Conjunto de términos conectados en el poliárbol de la red mixta.
$\mathcal{D}$	...	Conjunto de documentos de una colección.
d_j	...	Suceso correspondiente a "el documento D_j es relevante".
$\bar{d}_j$	...	Suceso correspondiente a "el documento D_j no es relevante".
D_j	...	Documento de una colección. Nodo documento en un grafo. Variable aleatoria binaria representando a un documento.
$D(P, P^T)$	...	Entropía de Kullback-Leibler de dos distribuciones de probabilidad.
$Dep(T_i, T_j)$	...	Entropía de Kullback-Leibler de dos variables término.
E	...	Conjunto de evidencias.
E_i^+	...	Conjunto de evidencias situado en el poliárbol de ancestros del nodo x_i.
E_i^-	...	Conjunto de evidencias situado en el poliárbol de descendientes del nodo x_i.
E_{u_j,x_i}^+	...	Evidencia que recibe x_i del padre u_j.
E_{x_i,y_j}^-	...	Evidencia que recibe x_i del hijo y_j.
f	...	Frecuencia de aparición de una palabra en un texto.
$\mathcal{F}$	...	Marco del modelo de recuperación.
FI_{D_j}	...	Fichero invertido para el documento D_j.
G	...	Grafo dirigido acíclico.
idf_i	...	Frecuencia documental inversa del término i-ésimo.
$I(X, Y \mid Z)$	...	Independencia condicional de X e Y dado Z.
$I(X, Y \mid \emptyset)$	...	Independencia marginal de X e Y.
$I(x_i, x_j)$	...	Medida de información mutua esperada entre dos variables.
i, j, k	...	Índices.

J	...	Conjunto de juicios de relevancia.
$\mid J \mid$	...	Cardinal de conjunto de juicios de relevancia.
m_j	...	Número de términos por los que ha sido indexado el documento D_j.
m'_j	...	Número de documentos padre que tienen el documento D_j en la subred de documentos extendida.
M	...	Número de términos de una colección.
$n(C)$	...	Número de documentos que incluyen los términos que están a relevantes en la configuración C y no aparecen los que son no relevantes en dicha configuración.
ndr	...	Número de documentos con los que relacionar a uno dado en el fichero invertido.
n_i	...	Número de documentos donde aparece el término i-ésimo de la colección.
n_i^R	...	Número de ocurrencias del i-ésimo término en R.
n_t	...	Número de ocasiones en el que el término T se ha observado entre los $\mid J \mid$ primeros documentos.
$n_{\bar{t}}$	...	Número de documentos de entre los $\mid J \mid$ primeros en los que no aparece el término T.
n_r	...	Número de documentos relevantes recuperados.
n_{rt}	...	Número de documentos relevantes recuperados que han sido indexados por T.
$n_{\bar{r}t}$	...	Número de documentos recuperados no relevantes indexados por T.
$n_{r\bar{t}}$	...	Número de documentos recuperados relevantes en los que no aparece T.
$n_{\bar{r}\bar{t}}$	...	Número de documentos recuperados no relevantes donde no aparece el término T.
N	...	Número de documentos de una colección.
ND	...	Nueva capa de documentos.
$Norm(w_{ij})$	...	Peso normalizado.
O_{D_j}	...	Ordenación decreciente de los documentos de la colección según la probabilidad de relevancia de cada uno dado D_j.
p	...	Posición de una palabra en la ordenación por frecuencias.
P	...	Distribución de probabilidad obtenida a partir de datos.
P^T	...	Distribución de probabilidad obtenida a partir de una estructura simplemente conectada.
qf_i	...	Peso del i-ésimo término en la consulta.
$\mathcal{Q}$	...	Conjunto de consultas.
Q	...	Consulta.
Q_{D_j}	...	Subconjunto de una consulta formado por aquellos términos que figuran en el documento D_j.

$\bar{r}$	...	Suceso correspondiente a ”el documento no es relevante”.
r	...	Suceso correspondiente a ”el documento es relevante”.
R	...	Conjunto de documentos relevantes a una consulta.
$R_{\pi(D_j)}$	...	Conjunto de términos que son relevantes en una configuración dada de los términos que indexan D_j.
$R_{\pi'(D_j)}$	...	Conjunto de documentos que son relevantes en una configuración dada de los documentos con los que está relacionado D_j.
$\mid R \mid$	...	Número de documentos relevantes a una consulta.
$\bar{R}$	...	Conjunto de documentos no relevantes a una consulta.
$\mid \bar{R} \mid$	...	Número de documentos no relevantes a una consulta.
RBD	...	Red bayesiana documental.
$s, s_t, s_{\bar{t}}$	...	Tamaños muestrales equivalentes en los estimadores bayesianos.
$\bar{S}$	...	Similitud media de los documentos de una colección.
$\bar{S}_i$	...	Similitud media de los documentos de una colección sin el término i-ésimo.
$S(D_i, D_j)$	...	Función que devuelve la similitud entre dos documentos.
$Sim(D_j, Q)$	...	Función que devuelve la similitud entre un documento y una consulta.
$tc+$	...	Conjunto de términos de la consulta positivos.
$tc-$	$\cdots$	Conjunto de términos de la consulta negativos.
$tc=$	$\cdots$	Conjunto de términos de la consulta neutros.
$te+$	...	Conjunto de términos de expansión positivos.
$te-$	$\cdots$	Conjunto de términos de expansión negativos.
$te=$	$\cdots$	Conjunto de términos de expansión neutros.
tf_{ij}	...	Frecuencia de aparición del término i-ésimo en el documento j-ésimo.
t_i	...	Suceso correspondiente a “el término t_i es relevante”.
$\bar{t}_i$	...	Suceso correspondiente a “el término t_i es relevante”.
$\mathcal{T}$	...	Conjunto de términos de una colección.
T_i	...	Término de la colección. Nodo término en un grafo. Variable aleatoria binaria representando a un término.
$\langle \mathbf{t_1}, \ldots, \mathbf{t_k} \rangle$	...	Configuración de valores de las variables $T_1, \ldots, T_k$.
$\mathcal{U}$	...	Conjunto de variables aleatorias.
vdt	...	Valor de discriminación de un término.
x, y, z, u	...	Variables aleatorias.
$\mathbf{x}, \mathbf{y}, \mathbf{z}$	...	Asignaciones de valores concretos a las variables.
$x \rightarrow y$	...	Arco en un G.D.A. desde el nodo x al nodo y.
X, Y, Z	...	Conjuntos de variables aleatorias.
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	...	Asignaciones de valores concretos a los conjuntos de variables.
$\langle X, Y \mid Z \rangle_G^d$	...	X está d-separado de Y por Z.
wd_{ij}	...	Peso del documento i-ésimo en el j-ésimo con el que está relacionado en la subred de documentos extendida.
w_{ij}	...	Peso del término i-ésimo en el documento j-ésimo.
w_{iQ}	...	Peso del término i-ésimo en la consulta.

$\lambda_i(x_i)$	...	Mensaje recibido por el nodo x_i de sus hijos.
$\lambda_{y_j,x_i}(y_j)$	...	Mensaje que el nodo x_i envía a su hijo y_j.
$\Pi(x)$	...	Conjunto de padres de un nodo x.
$\pi(x)$	...	Configuración para los padres del nodo x.
$\pi(D_j)$	...	Configuración para los padres término de un nodo documento D_j.
$\pi'(D_j)$	...	Configuración para los padres documento de un nodo documento D_j.
$(\pi(D_j))^{\downarrow T_i}$	...	Configuración de los padres de D_j excepto el término i-ésimo.
$\rho_i(x_i)$	...	Mensaje recibido por el nodo x_i de sus padres.
$\rho_{u_j,x_i}(u_j)$	...	Mensaje que x_i envía a su padre u_j.
ω_1	...	Suceso correspondiente a “un documento es relevante”.
ω_2	...	Suceso correspondiente a “un documento no es relevante”.

Introducción.

Actualmente, y debido al auge de la red Internet, existe una gran cantidad de información dispuesta en forma de texto, imágenes y sonido, distribuida a lo largo y ancho de la Red, con el inconveniente principal de que no se puede acceder de manera fácil y rápida. Además, existe el problema adicional de que la búsqueda de un material relevante a cierta necesidad de información, necesita de un gran esfuerzo para filtrar material que no es útil.

Estos son los problemas básicos que se tienen en nuestro tiempo a la hora de acceder a una información que nos puede resultar útil, pero no son problemas que han surgido en la actualidad, sino que aparecieron en el momento en el que el hombre decidió organizar de alguna forma todo el material del que disponía, dando lugar así a las bibliotecas y a todo el proceso que lleva asociado la gestión de sus fondos bibliográficos.

Fue por los años cuarenta del siglo pasado cuando realmente se constató la necesidad de establecer soluciones a ese problema y cuando se empezaron a presentar las primeras, apoyadas por la existencia de los primeros ordenadores [Ell97]. Así, la disciplina de la Recuperación de Información, que sería la encargada de establecer las técnicas para la organización y posterior acceso eficiente a la información [Rij79, SM83], comenzó su desarrollo, y su evolución se ha mantenido más o menos paralela a los avances tecnológicos en el mundo de la Informática.

La utilización de los ordenadores en la recuperación documental ha aportado una continua fuente de soluciones para organizar y acceder automáticamente a la información, dando lugar al desarrollo de multitud de modelos de recuperación [Rij79, SM83, FB92, Ell97, BR99]. A pesar de este aporte, el tratamiento por parte de la Informática de la información documental tiene todavía grandes inconvenientes que hacen que las soluciones ofertadas no sean todo lo buenas que se desearía. Uno de estos impedimentos radica en la incertidumbre intrínseca asociada a las diferentes tareas que componen esta disciplina, y que evita que los sistemas de recuperación de información le den al usuario exactamente aquellos documentos que satisfacen plenamente las necesidades que lo llevaron a realizar la consulta [TC97].

La Informática, en su rama de la Inteligencia Artificial, ha centrado parte de sus esfuerzos investigadores en aquellos problemas donde existen diferentes tonalidades de gris (no todo es blanco o negro) logrando a menudo considerables éxitos en el tratamiento de la incertidumbre, configurándose así como la única área posible donde se pueden resolver ciertos problemas.

Las redes de creencia [Pea88, Nea90, CGH96, Jen96], por su capacidad para alcanzar un gran rendimiento en la resolución de problemas con un alto grado de incertidumbre, se han conformado como un modelo de referencia básico en el campo de la Inteligencia Artificial al que pertenecen.

Así las cosas, los investigadores en el campo de la recuperación de información se percataron que el problema que tenían entre manos era idóneo para poner en práctica la robusta metodología aportada por las redes de creencia. Esta aplicación no fue directa, ya que se originó como una evolución natural de uno de los primeros modelos de recuperación desarrollados, el probabilístico [Rij79] basado, como indica su nombre, en la Teoría de la Probabilidad. La razón de esta evolución radica en que las redes de creencia, en una de sus versiones (la más extendida), se sustentan en el formalismo probabilístico, dando lugar entonces a lo que se conoce como Redes Bayesianas [Pea88].

La aplicación de las redes bayesianas a la recuperación de información tuvo su máximo apogeo a partir de los años noventa, cuando varios investigadores a lo largo de todo el mundo desarrollaron diferentes modelos, técnicas y, posteriormente, sistemas de recuperación de información [Tur90, TC91, TC91b, TC92, CT92, Bro94, Gre96, GCT97, CCH92, CCB95, GIS96, IGS96, RM96, Rei00, SRCMZ00, FCAT90, FF93, FF95, BI94, BG94b, Ijd94, IBH95, Sah96, Sah98, SDHH98, SYB98, DPHS98, FC89, SD91, Sav93, Sav95, TH93, CFH98, CFH00], demostrando empíricamente la idoneidad del uso de estas herramientas en este ámbito. Pero, a pesar del buen rendimiento ofrecido, existen peculiaridades del problema que el rendimiento no sea el máximo posible, dejando abierto un campo muy amplio a la investigación. En él es donde nosotros enmarcamos el trabajo que hemos desarrollado, ofreciendo un nuevo modelo de recuperación de información basado en redes bayesianas, con los objetivos principales de aumentar la expresividad con respecto a la ofrecida por los ya existentes, solucionar algunas carencias de los mismos y, consecuentemente, mejorar el rendimiento del sistema de recuperación de información basado en estas herramientas.

Planteamiento del problema.

Las hipótesis de trabajo de las que partimos son las siguientes:

- El incremento del número de documentos a los que podemos acceder hace necesario el desarrollo de herramientas automatizadas que nos permitan una representación, gestión y acceso eficiente a la información.
- Las redes bayesianas son formalismos de propósito general para la representación del conocimiento con incertidumbre, que pueden ser adaptadas para la resolución de una gran cantidad de problemas.
- La incertidumbre inherente al problema de la recuperación de información hace que el problema sea susceptible de ser abordado mediante el uso de redes bayesianas.

Por tanto, y fijándonos en estas suposiciones iniciales, podemos concluir que las redes bayesianas aportan características adecuadas para tratar la peculiaridad principal de la recuperación de información. Así, el problema que plantearemos en esta memoria será el desarrollo de modelos de recuperación de información que estén basados en las redes bayesianas, con objeto de alcanzar una capacidad de expresión y de recuperación altas.

La motivación principal que nos ha llevado a tratar este problema ha sido el estudio que se ha hecho de otros modelos ya existentes, viendo así la idoneidad de la herramienta sobre el problema que tratan y cómo éstos modelos poseen limitaciones y aspectos no completados que hacen que se puedan mejorar y afinar.

Objetivos.

El objetivo principal de esta memoria es el desarrollo de un modelo formal de recuperación de información basado en redes bayesianas y su posterior implantación en un sistema de recuperación totalmente operativo. Este objetivo genérico se puede descomponer en los siguientes, más específicos:

1. Establecer una representación en forma de red de creencia, basándonos en un formalismo probabilístico, de la información documental disponible, totalmente adaptada a las peculiaridades del problema objeto de estudio, de manera que se consiga un alto nivel de expresividad.

2. Adaptar los algoritmos de aprendizaje de redes bayesianas disponibles y, en su caso, diseñar nuevos métodos, con objeto de reducir los inconvenientes acarreados por las dimensiones del problema.

3. Modificar los algoritmos de propagación de probabilidades ya existentes y, también en su caso, desarrollar métodos alternativos, para llevar a cabo el proceso de recuperación de una manera eficiente y lo más exacta posible.

4. Determinar el rendimiento que se puede alcanzar con dicho modelo mediante su evaluación con una serie de colecciones documentales estándar. Comparar dicho rendimiento con otros modelos clásicos de recuperación de información, así como con los basados en redes de creencia.

5. Mejorar la calidad de recuperación del modelo mediante la incorporación de técnicas, como la realimentación de relevancia, totalmente adecuadas a las características del modelo desarrollado y que permitan incrementar el rendimiento del sistema.

Descripción de la memoria por capítulos.

Para alcanzar los objetivos indicados anteriormente, vamos a organizar esta memoria como describimos a continuación.

Capítulo 1. Este primer capítulo nos servirá para exponer los conceptos generales y básicos para comprender el resto de la memoria, tanto en la vertiente de las redes bayesianas como en la de recuperación de información. Comenzaremos introduciendo las bases de la Recuperación de Información mediante la descripción del proceso completo, desde la indexación de documentos y consultas, hasta las técnicas existentes para mejorar el rendimiento del sistema recuperador, pasando por la forma en que los principales modelos hacen la propia recuperación. La segunda parte se dedicará a hacer lo mismo, pero en este caso con las redes de creencia: qué son, para qué sirven, cómo se construyen (proceso de aprendizaje) y, por último, cómo se utilizan (propagación). Finalmente, uniremos estas dos áreas, pasando a centrarnos en una revisión de las principales aplicaciones de las redes de creencia a la recuperación de información.

Capítulo 2. El segundo capítulo pretende demostrar cómo las redes bayesianas pueden utilizarse eficiente y eficazmente como ayuda a cualquier sistema de recuperación de información, en este caso para realizar tareas de expansión de consultas. Describiremos detalladamente cómo construimos la red de creencia subyacente, es decir, explicaremos su topología y el algoritmo de aprendizaje diseñado a medida para este problema. Una vez hecho esto, pasaremos a comentar cómo se puede utilizar la red construida para poner en práctica la técnica de modificación de la consulta mediante expansión. Finalmente, concluiremos el capítulo con la experimentación efectuada con este sistema de apoyo a la recuperación.

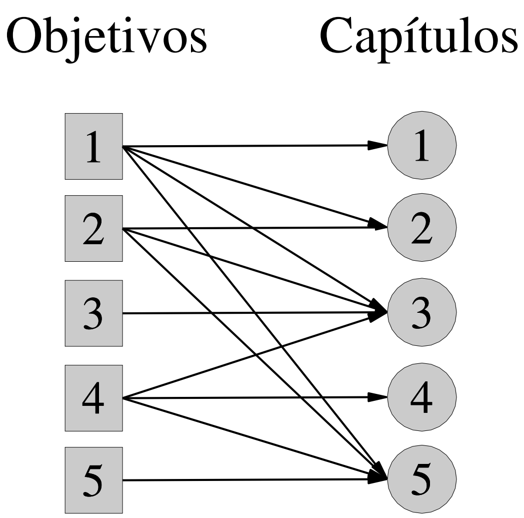
Capítulo 3. Este capítulo se puede calificar como el más importante de la memoria, ya que en él presentamos el modelo genérico de recuperación de información que se ha desarrollado. Se distinguen tres partes principales en él: la primera, que se centra en la descripción del modelo topología y estimación de las distribuciones de probabilidad; la segunda gira entorno a la exposición de cómo se puede usar el modelo construido para recuperar documentos, tarea que se pone en práctica mediante el correspondiente método de inferencia en redes bayesianas. Y, por último, la tercera parte centrada en la exposición de la experimentación y en el análisis de la misma.

Capítulo 4. Una de las técnicas más ampliamente utilizadas para mejorar el rendimiento de un sistema de recuperación de información es la conocida como realimentación de relevancia, por la cual la aplicación recuperadora genera nuevas consultas a partir de la información suministrada por el usuario, una vez que ha inspeccionado los documentos que le devuelve el sistema documental como respuesta a una consulta. Así, y tras realizar una introducción general a esta herramienta y otra específica, en la que se estudia cómo otros modelos basados en redes de creencia ponen en práctica esta técnica, pasaremos a describir dos propuestas novedosas para realizar la realimentación en nuestro modelo. Cada propuesta viene acompañada de una detallada experimentación, al mismo tiempo que de los correspondientes comentarios explicando los comportamientos.

Capítulo 5. Hasta este momento, se ha supuesto que los documentos incluidos en una colección no tienen relaciones directas unos con otros, sino solamente a través de los términos comunes que los indexan. En este capítulo presentaremos un mecanismo para relacionarlos entre sí, originando un modelo más expresivo. Así, y tras explicar los fundamentos de este método, mostraremos los resultados empíricos que se obtienen mediante el uso de esta extensión de los modelos originales.

Capítulo 6. Por último, expondremos las conclusiones obtenidas en el desarrollo de esta investigación y esbozaremos las líneas generales por las que discurrirá nuestra investigación en el futuro.

El siguiente gráfico muestra los capítulos de esta memoria en los que se abordan cada uno de los objetivos propuestos.



1. Introducción a la recuperación de información y a las redes bayesianas.

En este capítulo haremos un repaso de los conceptos básicos sobre recuperación de información (R.I.) y redes bayesianas, con objeto de tener el bagaje informativo necesario para seguir esta memoria. En primer lugar, nos centraremos en la recuperación de información, ofreciendo una visión general del proceso completo: desde la indexación de la colección de documentos, hasta finalmente la evaluación de la recuperación en sistemas experimentales. Se introducirá el concepto de modelo de recuperación y nos centraremos con algo más de detalle en los dos modelos de recuperación más relevantes para nuestro trabajo: el del espacio vectorial y el probabilístico. En cuanto a las redes bayesianas, después de introducirlas comentaremos cómo se usan, para lo cual se lleva a cabo el proceso de propagación de probabilidades y seguidamente cómo se construyen, proceso que se denomina aprendizaje. Por último, revisaremos los trabajos existentes que han aplicado las redes bayesianas a la recuperación de información, justificando el uso de estas herramientas en esta disciplina.

1.1. Conceptos generales sobre recuperación de información.

La disciplina de la R.I. surgió en los años cuarenta cuando la cantidad ingente de material científico exigía la creación de una herramienta que la gestionara [Fra92]. El problema, según van Rijsbergen [Rij79], es muy claro: existe una gran cantidad de información a la que se debe acceder de manera precisa y rápida. Una buena definición inicial que establece el proceso completo de R.I. la dan Salton y McGill en [SM83]: la R.I. trata de la representación, almacenamiento, organización y acceso de ítem de información.

Hablaremos de *recuperación de información automática* cuando todas las tareas anteriormente indicadas se lleven a cabo con un ordenador, definiendo un *sistema de recuperación de información (S.R.I.)* como el software que implementa estas tareas en un ordenador, con el objetivo fundamental de darle a un usuario que ha articulado una consulta toda la información que la satisfaga.

La definición inicial de R.I. ha tenido que ser actualizada debido al gran avance que se ha producido en los últimos años en esta disciplina, incluyendo áreas como el modelado, la clasificación de documentos, la categorización, arquitectura de sistemas, interfaces de usuario, visualización de datos, filtrado, etc. [BR99]. Pero ha sido con el auge de Internet y de la World Wide Web cuando los investigadores en esta área han ampliado los campos de investigación como consecuencia de los problemas que surgen al acceder al gran número de páginas web distribuidas por ordenadores de todo el mundo.

Otra cuestión a clarificar es el concepto de *ítem de información*, el cual inicialmente puede ser cualquier tipo de objeto, como por ejemplo una imagen, un sonido o, incluso, una escultura, pero que en nuestro caso lo interpretaremos como la representación textual de cualquier objeto y lo denominaremos genéricamente *documento*. Así, éste podrá ser un resumen de un artículo en una revista científica, el propio artículo, artículos en revistas o periódicos generales, libros, informes técnicos, descripciones de material audiovisual, páginas web, mensajes de correo electrónico o, incluso en ciertas ocasiones, capítulos, secciones o párrafos [SM83, Kor97].

Otro asunto importante es la forma en que se representará el documento en el ordenador. Por un lado, y teniendo en cuenta que la capacidad de almacenamiento externo de los ordenadores actuales ha crecido considerablemente, se podrían almacenar los documentos íntegramente, lo que se conoce como representación a texto completo. Pero hay casos en que, debido al gran tamaño de las colecciones, no es posible. Por tanto, hay que buscar una manera alternativa que pasa por el procesamiento del texto para obtener una representación del mismo en forma de *palabras clave o términos de indexación*: un conjunto de palabras que resume el contenido del documento. Con esta forma de proceder se evitan dos problemas: por un lado, se reduce el espacio físico necesario para almacenar la colección; y por otro, se eliminan palabras que no aportan información ninguna a la hora de describir el contenido del documento y que son fuentes de ruido para tareas futuras [BR99]. Otra tercera ventaja es que se expresan los documentos de una manera mucho más cómoda para que el ordenador los maneje eficientemente.

Pasemos seguidamente a describir cuál es el proceso completo de recuperación, representado gráficamente en la figura 1.1, y cuyas etapas individuales se introducirán más detalladamente en las secciones siguientes. Dada una colección de documentos, el primer paso, como ya hemos indicado, es obtener una representación textual de los mismos y, seguidamente, mediante la *indexación*, conseguir un conjunto de términos por cada documento. En este momento, la base de datos documental está lista para ser utilizada y es cuando interviene el usuario del S.R.I., formulándole una consulta que expresa su necesidad de información. El S.R.I. pone en marcha su motor de búsqueda y compara cada uno de los documentos almacenados con dicha consulta, obteniendo en algunos casos el grado con el que el documento satisface a la consulta, y en otros simplemente seleccionando sólo los documentos que la satisfagan completamente. El siguiente paso es presentar al usuario la salida del proceso de búsqueda, evaluará dicha salida y decidirá si ésta es satisfactoria o, por el contrario, no ha satisfecho completamente su necesidad de información. Este hecho originará que el S.R.I. vuelva a recuperar ayudado por la información adicional suministrada por el usuario, proceso que se conoce como *realimentación de relevancia* y que queda marcado por flechas discontinuas en el gráfico. En entornos experimentales, el S.R.I. incorpora un módulo adicional que se encargará de determinar la calidad de la recuperación del sistema, proceso denominado *evaluación de la recuperación*.

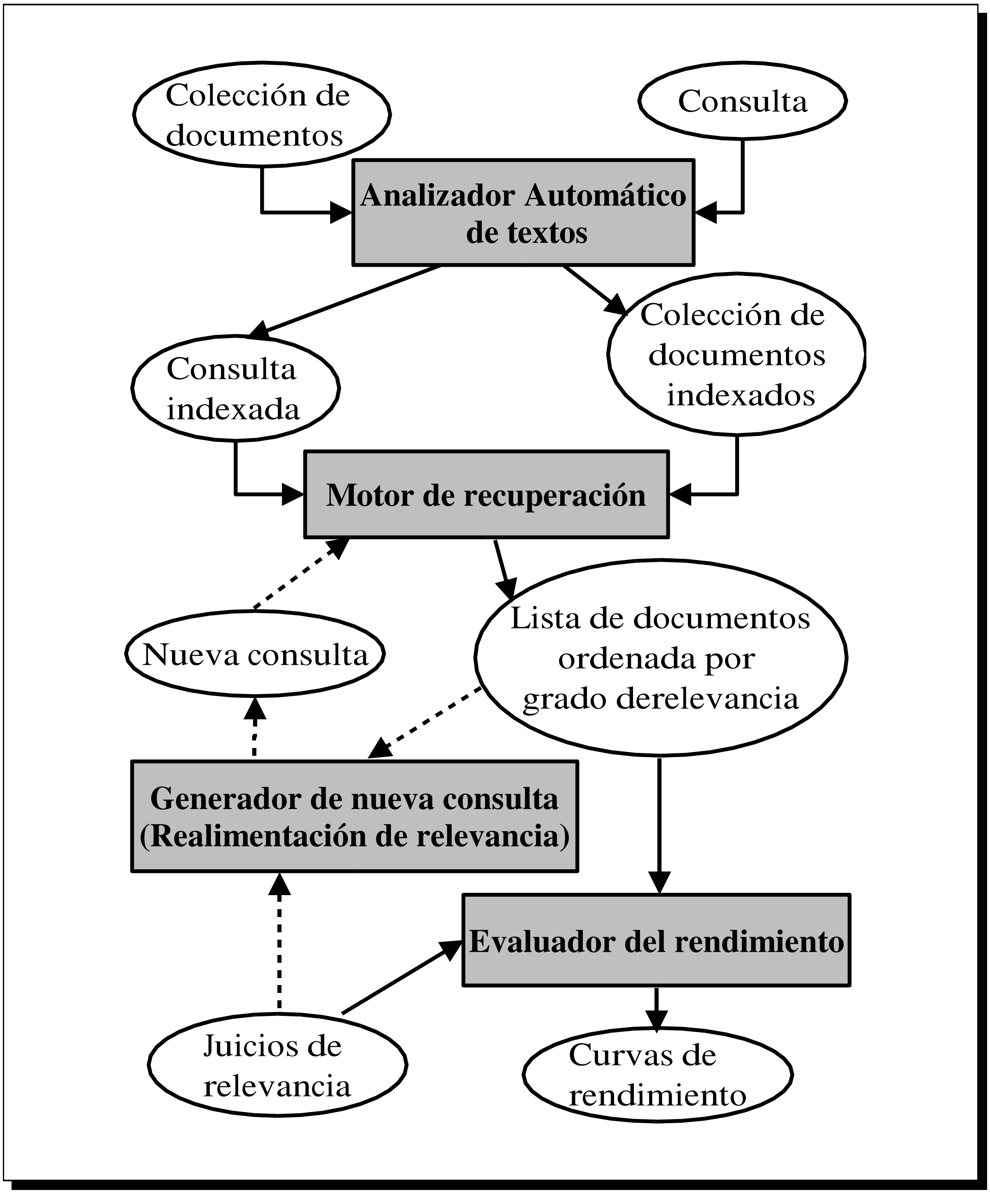


Figura 1.1.: Proceso completo de recuperación de información.

Un concepto importante en R.I. es el de *modelo de recuperación*, el cual de manera informal se puede definir como una especificación de la representación de los documentos y de las consultas, más la forma en que se compararán para recuperar los documentos relevantes. Baeza y Ribeiro presentan en [BR99] una caracterización formal de dicho concepto como una cuádrupla $(\mathcal{D}, \mathcal{Q}, \mathcal{F}, Sim(D_j, Q))$ donde:

- $\mathcal{D}$ es un conjunto compuesto por las representaciones para los documentos de la colección,
- $\mathcal{Q}$ es un conjunto formado por las representaciones de las necesidades de información de los usuarios,
- $\mathcal{F}$ es un marco para modelar representaciones de documentos, consultas y sus relaciones y
- $Sim(D_j, Q)$ es una función que asocia un número real al par (D_j, Q), $Q \in \mathcal{Q}$ y $D_j \in \mathcal{D}$, con el cual se puede realizar una ordenación de documentos con respecto a la consulta Q.

Relacionado de manera directa con la función *Sim*, una de las piedras angulares de la R.I. es el concepto de *relevancia*, que surgió paralelamente a esta disciplina sobre los años cuarenta [Sar75], aunque, y a pesar del tiempo transcurrido, actualmente sigue sin ser entendido plenamente y, por tanto, completamente definido [Miz97]. La definición que hacen Saracevic y col. en [SKCT88], a partir de la cual efectuaron sus experimentos, establece que un documento es relevante si el usuario considera que está relacionado plenamente con la consulta, y parcialmente relevante si opina que satisface de manera parcial dicha consulta (las afirmaciones sobre la relevancia o no de un documento expresadas por un usuario reciben el nombre de *juicios de relevancia*). O con otras palabras, el grado con el que un documento trata de la materia expresada en la consulta [SM83], considerándola así como un concepto multivaluado [Sar75], y no binario (relevante, no relevante). Como se puede observar en esta definición, es un criterio totalmente subjetivo del usuario, y así lo defiende Bookstein en [Boo79], estableciendo que un documento es relevante con respecto a un usuario si éste satisface la necesidad que lo llevó a examinar el documento en cuestión. De esta manera, dos usuarios distintos pueden decidir que un mismo documento es y no es, respectivamente, relevante a una consulta dada y, además, darle un grado diferente. Park en [Par97] introduce en el concepto de relevancia elementos adicionales como el propio S.R.I. y las representaciones de los documentos.

Así, el objetivo de la función *Sim* es determinar, en lugar del usuario, el grado de relevancia de un documento con respecto a una consulta. Esta tarea fue introducida por primera vez en el campo de la R.I. por Maron y Kuhns [MK60], los cuales dieron la forma de probabilidad a esa función, midiendo, por tanto, la probabilidad de que las similitudes entre un documento y una consulta condujeran a que el usuario aceptara ese documento como respuesta a dicha consulta.

El mismo Saracevic, en [Sar75], hizo un estudio muy detallado sobre la relevancia en R.I., al igual que Mizzaro posteriormente [Miz97]. Ambos revisan su evolución histórica, comentando las diferentes visiones que han caracterizado a este concepto a lo largo de los años. Por otro lado, Gluck, en [Glu96], también ofrece otra revisión de las diferentes definiciones de la relevancia, aunque bastante más somera, para posteriormente relacionarla con la satisfacción del usuario después de una recuperación.

1.1.1. Indexación.

El primer paso que debe dar un S.R.I. antes de estar totalmente preparado para que los usuarios lo utilicen es el de procesar la base de datos de documentos, dejándola en un formato cuya manipulación por parte del sistema sea fácil y rápida. Por tanto, a partir de un documento se generará una representación del mismo, formada por una secuencia de *términos de indexación*, los cuales mantendrán lo más fielmente posible el contenido original del documento. A este proceso general se le denomina *análisis automático de textos* [Rij79].

¿Qué términos son los que usaremos realmente para indexar un documento? La base para responder a esta pregunta, nos la da, por un lado, el trabajo que llevó a cabo Lunhs [Rij79, SM83], quien planteaba que la frecuencia de aparición de una palabra en un texto determina su importancia en él, sugiriendo que dichas frecuencias pueden ser utilizadas para extraer palabras con objeto de resumir el contenido de un documento. Por otro lado, está la ley de Zipf [Rij79, SM83], la cual establece que si obtenemos la frecuencia de aparición, f, de cada palabra de un texto y las ordenamos decrecientemente, siendo p la posición que ocupa en dicha ordenación, se cumple que $f \cdot p \simeq c$, donde c es una constante.

Si se representa gráficamente esta curva (p en el eje X, y f en el Y), se obtiene una hipérbola, en la cual se pueden establecer dos límites en cuanto a p se refiere (véase la figura 1.2): todas las palabras que excedan el superior, se considerarán muy comunes (haciendo una búsqueda por ellas podríamos recuperar casi todos los documentos), y todas las que estén por debajo del inferior, muy raras. Las que queden dentro de ambos límites serán las que tengan una mayor capacidad para discriminar el contenido de un texto y, por tanto, las que deban ser usadas. El problema radica en establecer los dos límites anteriores, porque, tal y como dicen Salton y McGill en [SM83], la eliminación de palabras con frecuencias muy altas puede provocar una reducción de la exhaustividad, ya que el uso de conceptos generales es útil a la hora de recuperar muchos documentos relevantes. Por el contrario, el descartar términos con una frecuencia baja, produce pérdidas en la precisión [1]. Intentando paliar estos problemas, Pao ofrece un método para calcular automáticamente el límite inferior [Pao78].

Otro aspecto a tener en cuenta a la hora de seleccionar los términos consiste en eliminar las palabras vacías de significado, como pueden ser artículos, preposiciones, conjunciones, incluso en algunos casos, se pueden calificar así algunos verbos, adverbios y adjetivos [BR99]. Por tanto, estas palabras no nos sirven como términos de indexación, ya que, por un lado son muy frecuentes, y por otro no representan correctamente el contenido del documento [Kor97]. La acción normal que se lleva a cabo con ellas es su eliminación del texto, proceso que se conoce como *eliminación de palabras vacías* (*stopwords* en inglés), y se pone en práctica mediante la

[1] Los conceptos de exhaustividad y precisión se tratarán detalladamente en la sección 1.1.3.

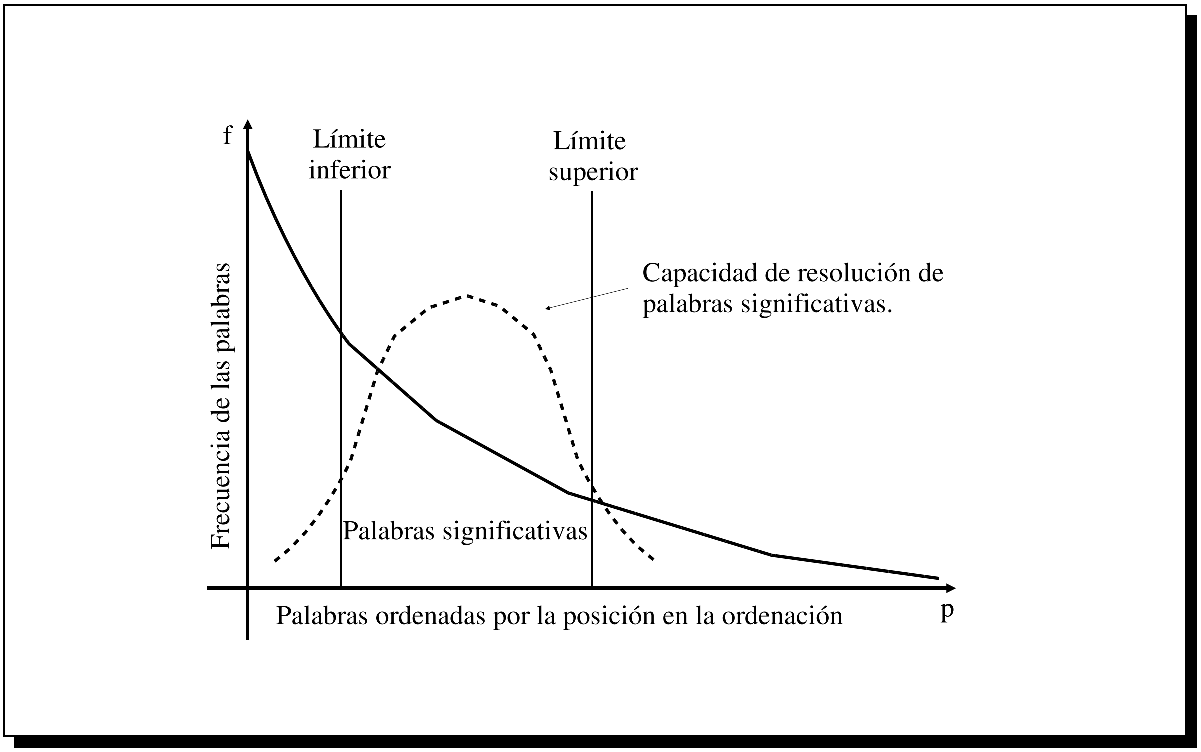


Figura 1.2.: Representación gráfica de la frecuencia de los términos ordenados según su posición en la ordenación: ley de Zipf.

comparación de cada palabra del texto con un diccionario que contiene la lista de palabras no aptas para la indexación (tanto en [Rij79] como en [Fox92] se presentan dos listas completas de palabras vacías).

Este proceso de selección pasa por determinar la importancia de un término en el documento, de tal forma que, si es lo "suficientemente" importante, se escogerá para ser incluido en el conjunto de términos final. El cálculo de la importancia de cada término se conoce como *ponderación del término*.

¿Cómo se mide esa importancia? Un primer enfoque se basa en contar las ocurrencias de cada término en un documento, medida que se denomina *frecuencia del término i-ésimo en el documento j-ésimo*, y se nota como $tf_{i,j}$. El problema que puede aparecer es que, independientemente del número de veces que aparezca el término "ordenador" en los documentos de una colección sobre Informática, no sería un buen término para asignárselo a ellos.

Una segunda medida de la importancia del término es la conocida como *frecuencia documental inversa* de un término en la colección, conocida normalmente por sus siglas en inglés: *idf (inverse document frequency)*, y que responde a la siguiente expresión [BR99]:

$$idf_i = \log \frac{N}{n_i} + 1 \tag{1.1}$$

Donde N es el número de documentos de la colección, y n_i el número de documentos donde se menciona al término i-ésimo. Como se puede observar, el valor obtenido por la expresión (1.1) decrece conforme n_i crece, variando desde $\log N + 1$ cuando n_i es 1, a 1 cuando n_i toma el valor N. Por tanto, cuantas menos veces aparezca un término en la colección, más alto será su idf [Kor97], dando así una forma de medir la calidad global del término en toda la colección. El hecho de introducir un logaritmo se justifica para suavizar el crecimiento del tamaño de la colección.

Lo ideal sería combinar ambas medidas anteriores utilizando un esquema de ponderación que permita identificar a los términos que aparecen bastante en varios documentos individuales, y a la vez, que se hayan observado en contadas ocasiones en la colección completa. Estos son los términos que tendrán una capacidad de discriminación mayor con respecto a los documentos en los que aparecen. O lo que es lo mismo, calcular un peso que fuera proporcional a la frecuencia del término i-ésimo en el documento j-ésimo, e inversamente proporcional al número de documentos de la colección completa en los que aparece ese término. Así, el peso final asignado al término i-ésimo en el documento j-ésimo, que notaremos como $tf \cdot idf$, corresponde al producto:

$$tf_{ij} \cdot idf_i \tag{1.2}$$

En este caso, la importancia crece con respecto a la frecuencia del término en el documento y disminuye con respecto al número de documentos que lo contienen [Kor97]. Cuanto más alto sea el valor, mejor será el término desde el punto de vista de la indexación.

Existen otras medidas como son el *valor de discriminación del término*, y la *relación señal/ruido* [SM83, Kor97], que se plantean como alternativas totalmente viables al $tf \cdot idf$, al igual que otra basada en la distribución estadística de Poisson [Har75a], la cual dio pie al desarrollo de un modelo de indexación completo [Har75b].

A partir de este momento, vamos a notar el conjunto de documentos de una colección ya indexado como $\mathcal{D}$, el glosario de la misma como $\mathcal{T}$ y sus tamaños como N y M, respectivamente. Un documento concreto se representará por el vector D_j:

$$D_j = ((T_1, w_{1j}), ((T_2, w_{2j}), \ldots, ((T_M, w_{Mj})) \tag{1.3}$$

donde $T_1, T_1, \ldots, T_M \in \mathcal{T}$, son los M términos del glosario de la colección y $w_{1j}, w_{2,j}, \ldots, w_{M,j}$ son los pesos respectivos de cada uno de éstos términos en el documento j-ésimo. Si un término no aparece en un documento, su peso será cero. En un S.R.I. real, sólo se almacenan los términos del documento por los que ha sido indexado, representándose el documento de la siguiente forma:

$$D_j = ((T_{1,j}, w_{1j}), ((T_{2j}, w_{2j}), \ldots, ((T_{m_j j}, w_{m_j,j})) \tag{1.4}$$

Así, T_{ij} es el i-ésimo término del j-ésimo documento y m_j es el número de términos por los que ha sido indexado el documento j-ésimo.

El último paso es extraer la raíz morfológica de cada palabra, eliminando sufijos y prefijos, originando así que el S.R.I. pueda recuperar documentos incluyendo variantes morfológicas de los términos contenidos en la consulta [Fra92b], mejorando la recuperación, a la vez de ahorrar espacio al almacenar sólo las raíces [SM83, BR99]. De entre los numerosos métodos para extraer las raíces [Rij79, SM83, Fra92b], hay que destacar por su simplicidad y efectividad el diseñado por Porter [Por80]. Por último, se debe poner en práctica un proceso de reconocimiento de raíces equivalentes [Rij79], con objeto de evitar confusiones con palabras que poseen la misma raíz pero no están relacionadas en su significado. A partir de este punto, al hablar de términos de indexación, nos estaremos refiriendo a las raíces morfológicas, en lugar de a las palabras completas.

Una vez que ha finalizado el análisis automático de la base de datos documental, un aspecto importante es su organización para conseguir un acceso eficiente y rápido en las operaciones que se realizarán posteriormente en el proceso de recuperación. Así, se conoce como *fichero invertido* a una estructura de datos que almacena de manera ordenada todos y cada uno de los términos del glosario y, para cada uno de ellos, guarda la lista de documentos donde aparece, junto con su peso asociado [HFBL92].

Cuando se efectúa una consulta al S.R.I., ésta es pasada también por el módulo de indexación para conseguir su correspondiente representación. Dependiendo del modelo de recuperación utilizado, la consulta podrá ser una expresión booleana, formada por los términos y conectivos lógicos, o una lista de términos, con sus correspondientes pesos, al estilo de los documentos. La consulta se notará como Q, y tendrá la siguiente forma:

$$Q = ((T_{1,Q}, w_{1,Q}), ((T_{2,Q}, w_{2,Q}), \ldots, ((T_{M,Q}, w_{M,Q})) \tag{1.5}$$

Igualmente, cuando un término no aparece en la consulta, su peso será cero, siendo la representación de la consulta análoga a la del documento.

1.1.2. Modelos de recuperación.

Existe una gran cantidad de modelos de recuperación basados en tecnologías muy diferentes. Belkin y Croft ofrecen en [BC87] una clasificación de las principales técnicas de recuperación y Baeza y Ribeiro hacen lo propio mediante una revisión de los diferentes modelos existentes actualmente [BR99]. En esta subsección vamos a tratar brevemente los tres clásicos: el booleano, el del espacio vectorial y el probabilístico. Además, esbozaremos otros modelos avanzados relacionados con la Inteligencia Artificial como son el basado en la teoría de subconjuntos difusos y en las redes neuronales.

1.1.2.1. El modelo booleano.

El modelo booleano está basado en la teoría de conjuntos y en el álgebra booleana. Su marco está compuesto por los documentos representados como conjuntos, las consultas, como expresiones booleanas (términos conectados por los conectivos booleanos *Y, O, y NO*), y las operaciones existentes para tratar conjuntos: unión, intersección y complementario [BR99, War92]. Los pesos de los términos en los documentos son binarios: 0 indica ausencia, y 1 presencia. Así, en este modelo, dada una consulta al sistema, se va evaluando la expresión booleana mediante la realización de las operaciones anteriores con los conjuntos formados por los documentos donde aparece cada término de la consulta (obtenido del fichero invertido). El conjunto de documentos resultante está compuesto por todos aquéllos que hacen verdad la consulta booleana: la función $Sim(D_j, Q)$ devolverá uno, si el documento j-ésimo hace verdad la expresión booleana Q, y cero, en caso contrario. Es por esto que Belkin y Croft en su clasificación lo enmarcan dentro de los modelos de emparejamiento exacto.

La ventaja principal de este modelo es su simplicidad, aunque existen varios inconvenientes a su uso, como pueden ser que los documentos son relevantes o no relevantes, descartando cualquier gradación de la relevancia, que la formulación de consultas booleanas puede llegar a ser una tarea algo compleja, y otras que se plantean en [Coo88, BC87], algunas de ellas con sus respectivas soluciones.

1.1.2.2. El modelo del espacio vectorial.

En el modelo del espacio vectorial [SL68, Sal71, SWY75] el marco está compuesto por el espacio vectorial de dimensión M (cada dimensión equivale a un término distinto del glosario), representando en él los documentos, las consultas y las operaciones algebraicas sobre los vectores de dicho espacio. Concretamente, la función que obtiene la similitud de un documento con respecto a una consulta se basa en la medida del coseno [SM83], la cual devuelve el coseno del ángulo que forman ambos vectores en el espacio vectorial. Esta medida tiene la siguiente forma:

$$Sim(D_j, Q) = \frac{\sum_{k=1}^{M} w_{kj} \cdot w_{kQ}}{\sqrt{\sum_{k=1}^{M} w_{kj}^2}\sqrt{\sum_{k=1}^{M} w_{kQ}^2}} \tag{1.6}$$

Así, cuando ambos vectores son exactamente el mismo, el ángulo que forman es de cero grados y su coseno es uno. Por el contrario, cuando el ángulo es de noventa grados (los vectores no coinciden en ningún término), el coseno será cero. El resto de posibilidades indicarán una correspondencia parcial entre el vector documento y el consulta, ofreciendo así una gradación en los valores de relevancia, de modo que cuanto más cercanos sean los vectores del espacio, más similares serán éstos..

En la fórmula anterior, el denominador tiene como función la de normalizar el resultado, mediante el producto de las longitudes euclídeas de los vectores. La primera, la del documento, tiende a penalizar los documentos con muchos término. Por el contrario, la segunda longitud, la de la consulta, no afectan la ordenación final de documentos.

1.1.2.3. El modelo probabilístico.

El marco del modelo probabilístico está compuesto por conjuntos de variables, operaciones con probabilidades y el teorema de Bayes.

Todos los modelos de recuperación probabilísticos están basados en el que hemos traducido como el *Principio de la ordenación por probabilidad*, conocido originalmente como *"the probability ranking principle"*. Este principio, formulado por Robertson en [Rob77], asegura que el rendimiento óptimo de la recuperación se consigue ordenando los documentos según sus probabilidades de ser juzgados relevantes con respecto a una consulta, siendo estas probabilidades calculadas de la forma más precisa posible a partir de la información disponible. Así, y atendiendo a este principio, el objetivo primordial de cualquier modelo probabilístico, pasa por calcular $p(R \mid Q, D_i)$, es decir, la probabilidad de relevancia dados una consulta y un documento.

Comencemos esta introducción a los modelos probabilísticos por el primero que surgió, el conocido como *modelo de recuperación con independencia binaria*, en inglés *Binary Independence Retrieval (B.I.R)*, que fue inicialmente planteado por Maron y Kuhns en [MK60], continuado por Robertson y Spark Jones [RJ76] y concluido por van Rijsbergen en [Rij79].

En él, los documentos y las consultas se representan por un vector binario. Así, un documento cualquiera tiene la siguiente forma:

$$D_j = (T_1, T_2, \ldots, T_M)$$

donde $T_i = 0$ ó 1 indica la ausencia o presencia del término i-ésimo, respectivamente, y M el número de términos de la colección. Existen dos eventos mutuamente excluyentes: ω_1, que representa el hecho de que un documento sea relevante, y ω_2, que indica que no lo sea. Este modelo asume que se conocen, o por lo menos se suponen, el conjunto de documentos relevantes (R) y no relevantes ($\bar{R}$) de una consulta dada.

El objetivo que se persigue es calcular $p(\omega_1 \mid D_j)$ y $p(\omega_2 \mid D_j)$, es decir, la probabilidad de que el documento D_j sea relevante y no relevante, respectivamente, dada una consulta Q y desarrollar una función que ofrezca un valor de relevancia para así poder ordenar los documentos según ella. En este caso, esa función tendrá la forma:

$$Sim(D_j, Q) = \frac{p(\omega_1 \mid D_j)}{p(\omega_2 \mid D_j)} \tag{1.7}$$

Haciendo suposiciones de independencia entre términos y aplicando el teorema de Bayes, se llega a:

$$Sim(D_j, Q) \sim \sum_{i=1}^{M} \log \left(\frac{p(T_i = 1 \mid \omega_1) \cdot (1 - p(T_i = 1 \mid \omega_2))}{p(T_i = 1 \mid \omega_2) \cdot (1 - p(T_i = 1 \mid \omega_1))} \right) T_i + c, \tag{1.8}$$

donde

	Relevante	No Relevante	
Aparece	n_i^R	$n_i - n_i^R$	n_i
No aparece	$\mid R \mid - n_i^R$	$N - n_i - \mid R \mid + n_i^R$	$N - n_i$
	$\mid R \mid$	$N - \mid R \mid$	N

Cuadro 1.1.: Distribución de la aparición o no de un término en los documentos relevantes y no relevantes.

$$c = \sum_{i=1}^{M} \log \frac{1 - p(T_i = 1 \mid \omega_1)}{1 - p(T_i = 1 \mid \omega_2)}, \tag{1.9}$$

siendo $p(T_i = 1 \mid \omega_1)$ la probabilidad de que un término T_i esté presente en el conjunto de documentos relevantes y $p(T_i = 1 \mid \omega_2)$ en los no relevantes. El logaritmo que multiplica al peso binario T_i, en la expresión (1.8), se conoce como el *peso de relevancia del término*: el valor que se le asigna a cada término cuando se está llevando a cabo una indexación probabilística, expresando la capacidad de discriminación de éste entre documentos relevantes y no relevantes.

La tabla 1.1 representa una tabla de contingencia para un término de la colección y muestra la distribución de apariciones o no del término i-ésimo en los documentos relevantes y no relevantes para una consulta. Dado que R es el conjunto de documentos relevantes, y $\mid R \mid$ su cardinal, N es el número total de documentos de la colección, n_i es el número de documentos en los que aparece T_i y n_i^R es el número de veces que aparece el término en documentos relevantes, las probabilidades $p(T_i = 1 \mid \omega_1)$ y $p(T_i = 1 \mid \omega_2)$ se estiman según las siguientes expresiones:

$$p(x_i = 1 \mid \omega_1) = \frac{n_i^R}{\mid R \mid} \text{ y } p(T_i = 1 \mid \omega_2) = \frac{N - n_i^R}{N - \mid R \mid} \tag{1.10}$$

El uso del modelo probabilístico que se acaba de presentar es el siguiente: el usuario formula una consulta al S.R.I. y éste, mediante la expresión (1.8), calcula un valor de relevancia para cada documento, generando así una lista ordenada de documentos. Cuando el usuario ha formulado una primera consulta, el S.R.I. no tiene información para poder estimar $p(T_i = 1 \mid \omega_1)$ y $p(T_i = 1 \mid \omega_2)$, según las expresiones (1.10), por lo que se deben establecer estimaciones iniciales, a partir de la colección completa, que pueden ser [BR99]:

$$p(T_i = 1 \mid \omega_1) = 0{,}5 \text{ y } p(T_i = 1 \mid \omega_2) = \frac{n_i}{N} \tag{1.11}$$

Croft y Harper ofrecen, en [CH79], varias estimaciones iniciales para cuando no hay información relevante y los rendimientos alcanzados con cada una de ellas. Por otro lado, Spark Jones, en [Jon79], establece varias expresiones cuando la información de la que se dispone es muy poca para obtener las tablas de contingencia de cada término.

A partir de la primera lista de documentos, el usuario emite sus juicios de relevancia con respecto a los documentos que figuran en ella y el S.R.I. genera la tabla 1.1, donde sí podrá aplicar directamente las expresiones (1.10) y reiterar este proceso hasta que el usuario quede satisfecho.

Existen otros modelos probabilísticos que surgieron como variación o mejora de este anterior. Entre ellos podemos destacar el conocido como *modelo de indexación de independencia binaria (Binary Independence Indexing Model, B.I.I.)* [FB91], que se desarrolló a partir del de Maron y Kuhns. Mientras el modelo de recuperación de independencia binaria trabaja con los documentos de la colección y una consulta, este modelo trabaja con un conjunto de consultas y el peso de cada término lo calcula con respecto a las consultas que usan ese término. Otros modelos probabilísticos importantes son los conocidos como *el enfoque de indexación Darmstadt* (Darmstadt Indexing Approach (DIA)) [Fuh89] o el modelo de *recuperación con indexación probabilística* [Fuh89, CLRC98] (*Retrieval with Probabilistic Indexing (R.P.I.)*. En [RRP80, Cro81, Boo83, Fuh89, Fuh92, FP94, Sav95, CLRC98] el lector podrá encontrar un estudio detallado de todos éstos y algunos otros que no citamos en esta memoria.

1.1.2.4. Otros modelos de recuperación.

Tratando inicialmente de resolver las limitaciones del modelo booleano, se aplicó *Teoría de subconjuntos difusos* desarrollada por Zadeh [Zad65] a la R.I., surgiendo así lo que se conoce como *recuperación de información difusa* [Boo80, Cro94, KBP99].

La recuperación difusa se centra en dos grandes líneas de investigación [KBP99]. Por un lado, están las extensiones al modelo booleano, dando lugar a los *modelos booleanos difusos extendidos*. Trabajan en la mejora de las representaciones de los documentos (se representan como subconjuntos difusos de términos cuyos pesos equivalen a los grados de pertenencia del término) y en la construcción de consultas más expresivas, (de nuevo añadiendo pesos asociados al valor devuelto por las funciones de pertenencia y generalizando los operadores booleanos clásicos). Por otro, cabe destacar lo que se conoce como *mecanismos asociativos difusos*, basados en el desarrollo y posterior uso de tesauros y pseudotesauros difusos, generalmente mediante técnicas de clasificación difusa.

Otro modelo de recuperación basado en una técnica de Inteligencia Artificial es el sustentado por las *Redes Neuronales*. El modelo representa los documentos mediante patrones de nodos (que corresponden a los términos) y conexiones ponderadas. Una consulta, por el contrario, se configura como el patrón de entrada a la red y la recuperación se hace activando la red para un entrada dada y encontrando los patrones almacenados a partir de dicha entrada [MCKH90, ORM91, Sch93].

1.1.3. Evaluación de la recuperación.

Un S.R.I. comercial, una vez que ordena los documentos según la similitud a la consulta y entrega esta lista al usuario, habría acabado su función (salvo que se desee aplicar algún tipo de método para mejorar la consulta), pero si se está utilizando un sistema experimental, se pondrá en marcha un último módulo que determinará el rendimiento del mismo. Este rendimiento puede calcularse en función del elemento que se evalúe: desde el tiempo que tardaría el sistema en responder después de efectuar una consulta, hasta la cantidad de material relevante que devuelve para una consulta dada o la cantidad de material devuelto que es relevante, pasando por cuestiones como la forma de presentar las salidas, el espacio que necesita para llevar a cabo las acciones correspondientes, o el esfuerzo requerido por el usuario para formular la consulta [Rij79, SM83, BR99].

Nos vamos a centrar seguidamente, y de forma exclusiva, en la evaluación de cómo de preciso es a la hora de recuperar el material relevante el S.R.I. Esta tarea se puede llevar a cabo mediante el cálculo de dos medidas: la *exhaustividad* (*recall* en inglés), y la *precisión*, definidas como sigue [SL68]:

$$\text{Exhaustividad} = \frac{\text{Número de documentos relevantes recuperados}}{\text{Número de documentos relevantes}} \tag{1.12}$$

$$\text{Precisión} = \frac{\text{Número de documentos relevantes recuperados}}{\text{Número de documentos recuperados}} \tag{1.13}$$

Es decir, la proporción de material relevante recuperado y la proporción de material recuperado que es relevante, respectivamente. Generalmente, la exhaustividad se incrementa cuando el número de documentos recuperados también lo hace y, al mismo tiempo, la precisión decrece [SM83] pero, como se puede observar de las expresiones anteriores, su rango quedará limitado por el intervalo $[0{,}0, 1{,}0]$. Habrá usuarios que estén interesados en niveles altos de exhaustividad, por lo que efectuarán consultas muy generales, y otros que lo estén en niveles altos de precisión, diseñando consultas muy específicas. De todas formas, es altamente deseable que la exhaustividad se acerque lo más posible al 100 %, al igual que la precisión, aunque ambos objetivos no se pueden alcanzar a la vez porque están inversamente relacionadas, como demuestran empíricamente Buckland y Key en [BG94]. Por otro lado, parece que la exhaustividad es más importante para los usuarios que la precisión [Su94].

El problema fundamental que poseen estas dos medidas es que la precisión se calcula de manera exacta, mientras que la exhaustividad no, ya que no se tiene un conocimiento claro de cuántos documentos relevantes existen en una colección para una consulta dada, por lo que se tendrá que estimar. Existen ocasiones en las que sí se conocen esos documentos relevantes: cuando la colección es muy pequeña y se pueden revisar uno por uno todos los documentos que la componen, y también en el caso de trabajar con colecciones de prueba en las que hay una batería de consultas con sus correspondientes documentos juzgados como relevantes por quien las efectuó [Kor97].

Supongamos ahora que se conocen cuáles son los documentos relevantes para una consulta dada, y también que disponemos de la lista de documentos ordenada por su grado de relevancia. Las medidas de exhaustividad y precisión van a cambiar conforme vaya variando (aumentando) el número de documentos recuperados, de acuerdo con la ordenación por relevancia. Una vez que se tienen todas las medias, se puede representar gráficamente ese conjunto de puntos (la exhaustividad en el eje X y la precisión en el Y). Generalmente, se puede observar cómo la precisión decrece conforme la exhaustividad aumenta [WMB99]. Con estas curvas, por ejemplo, se puede comparar el rendimiento para dos técnicas diferentes aplicadas sobre la misma consulta. Para ello, se suele llevar a cabo una interpolación en once valores estándar de exhaustividad: de 0,0 a 1,0, con incrementos de una décima. La descripción detallada del método de interpolación se encuentra en cualquiera de las referencias [SL68, Rij79, SM83, Kor97, BR99].

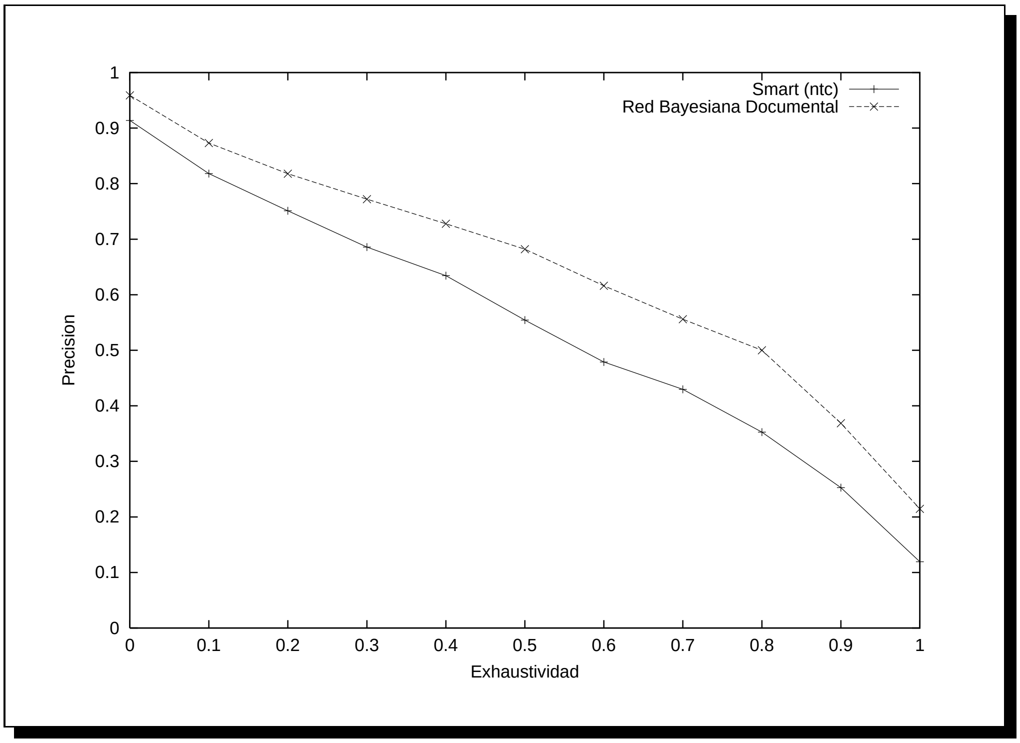


Figura 1.3.: Curvas E-P correspondientes a dos experimentos sobre la misma colección.

En el caso de que se desee establecer una única curva que resuma el comportamiento del S.R.I. sobre una batería de consultas, se procede a hacer la media de la precisión para cada uno de los once puntos de exhaustividad para cada consulta. Además, y para conseguir una única medida del rendimiento, se suele calcular la precisión media de los once puntos de exhaustividad para una curva dada, por un lado, y para los tres intermedios (0,2, 0,5 y 0,8) por otro.

Para comparar dos curvas, una de las técnicas más empleadas es el cálculo del porcentaje de cambio de la precisión media de una con respecto a otra: una de las curvas se establece como referencia, y a partir de ella se determina cuánto ha mejorado o empeorado la otra. Como ejemplo, en la figura 1.3 se representan las curvas obtenidas tras efectuar una batería de 30 consultas mediante dos S.R.I. diferentes sobre la colección MEDLARS. Vemos cómo es muy fácil, mediante esta gráfica, comparar el comportamiento de varios métodos de recuperación. El modelo cuya curva está etiquetada como “Red Bayesiana Documental” es considerablemente superior que el etiquetado como “SMART”, ya que para los mismos valores de exhaustividad,

	SMART	EXP1
REL. REC.	260	287
EXHAUST.	PRECISIÓN	
0.00	0.9137	0.9590
0.10	0.8181	0.8732
0.20	0.7512	0.8179
0.30	0.6858	0.7721
0.40	0.6344	0.7278
0.50	0.5543	0.6819
0.60	0.4789	0.6162
0.70	0.4297	0.5560
0.80	0.3527	0.4999
0.90	0.2527	0.3686
1.00	0.1193	0.2145
M. 11PTS	0.5446	0.6443
%C. 11PTS		18.3
M. 3PTS	0.5527	0.6666
%C. 3PTS		20.6
E. EX.	0.4104	0.4503
P. EX.	0.5778	0.6378
Exp1 → Red Bayesiana Documental		

Cuadro 1.2.: Curvas E-P comparando dos modelos de recuperación.

todos los de precisión son superiores en la primera.

De manera complementaria a las curvas de exhaustividad y precisión, se suele presentar la información cuantitativa obtenida como resultado del proceso de evaluación. La tabla 1.2 es un ejemplo, mostrando a la vez la estructura de las tablas de evaluación que vamos a seguir en esta memoria para exponer nuestros resultados.

Esta tabla ofrece el número de documentos relevantes recuperados (REL. REC) por los dos métodos que comparamos. Seguidamente, los valores de precisión (PRECISIÓN) obtenidos para los once valores estándar de exhaustividad (EXHAUST.). La etiqueta "M. 11PTS" se refiere a la media de los once puntos y " %C. 11PTS" al porcentaje de cambio que se produce al aplicar el segundo método con respecto al primero según las medias de precisión correspondientes a los once puntos. "M. 3PTS" se refiere ahora a la media de los valores de precisión correspondientes a los valores 0,2, 0,5 y 0,8 de exhaustividad y su correspondiente porcentaje de cambio. Por último, se ofrece la exahustividad y la precisión exactas (E. EX. y P. EX., respectivamente), es decir, la exahustividad y la precisión para el conjunto de documentos recuperados. En este caso, se han considerado los quince documentos mejores situados en la ordenación.

1.1.4. Métodos para mejorar la recuperación.

Hay veces en que el usuario no es capaz de encontrar una consulta que exprese fielmente su necesidad de información u ocasiones en las que, por limitaciones del propio S.R.I., éste no consigue recuperar todos los documentos relevantes. Por estas y otras razones se han desarrollado técnicas que permiten asistir al usuario a la hora de formular la consulta, por un lado, y por otro, reformular la consulta de manera iterativa a la luz de los juicios de relevancia expresados por el usuario.

Aunque en los capítulos 2 y 4 entraremos con algo más detalle sobre las técnicas existentes para mejorar la recuperación, en esta subsección las vamos a esbozar brevemente.

Tesauros. En primer lugar, y en cuanto a que ayuda al usuario para formular la consulta, destacamos el uso de los tesauros: un conjunto de palabras y las relaciones que existen entre sí, las cuales van desde sinonimias y antonimias hasta cualquier otro tipo de relación entre ellas. El tesauro puede usarlo el propio usuario para expresar su necesidad de información, mediante la búsqueda de las palabras adecuadas o, alternativamente, se suele utilizar como fuente para añadir nuevos términos a una consulta, proceso que se conoce como *expansión de consultas*.

Realimentación de relevancia. Por otro lado, siguiendo el estilo de trabajo que tiene el modelo probabilístico, existe la técnica conocida como *realimentación de relevancia*, la cual parte de un conjunto de juicios de relevancia expresados por el usuario tras una primera recuperación. Con esa información suministrada al S.R.I., éste modifica la consulta de dos maneras:

- Alterando los pesos de los términos que componen la consulta original, de tal forma que se le da más fuerza a aquéllos que aparecen más en documentos relevantes que en no relevantes, y debilitando a los que ocurren en la situación contraria. Esta técnica se conoce como *repesado de los términos*.

- Añadiendo nuevos términos que no aparecen en la consulta original, pero que sí lo hacen en los documentos que han sido juzgados como relevantes o no relevantes. Este proceso es también conocido como *expansión de consultas*.

1.1.5. Introducción al S.R.I. SMART.

Basado en el modelo del espacio vectorial, el S.R.I. *SMART* [SL68, Sal71, SM83, Buc85] ha sido y sigue siendo referencia para multitud de investigadores en todo el mundo. Fue desarrollado en los años sesenta en la Universidad de Cornell por un grupo de investigadores dirigidos por G. Salton y está disponible en la dirección *ftp.cs.cornell.edu* del Departamento de Ciencias de la Computación de la citada universidad.

SMART está formado por un conjunto de programas que componen un sistema completo de recuperación automática de documentos. Permite la creación, mantenimiento y uso de colecciones de documentos, de tamaños pequeños a medios. Pero ante todo, sus autores lo definen como una herramienta experimental para investigar métodos y técnicas de R.I. Además de crear una herramienta flexible, apta para la experimentación, los desarrolladores no se olvidaron de los usuarios, por lo que elaboraron un entorno rápido, portable e interactivo (a pesar de no poseer una interfaz de usuario gráfica).

Este S.R.I. está compuesto por cuatro módulos básicos:

- Módulo de indexación: convierte cualquier colección de documentos en su formato original a vectores de términos.
- Módulo de recuperación: calcula la similitud, basada en la función del coseno (expresión (1.6)), entre los documentos y una consulta, ya indexados previamente, generando como resultado una lista ordenada decrecientemente por dicha similitud de todos los documentos de la colección que es mostrada al usuario. Además, implanta una organización de los documentos con contenidos próximos en grupos, facilitando así la posterior recuperación.
- Módulo de realimentación de relevancia: a partir de los resultados de una consulta ya formulada y de los juicios de relevancia expresados por el usuario, genera una nueva consulta para recuperar más documentos relevantes.
- Módulo de evaluación, a partir de dicha lista ordenada y de los juicios de relevancia establecidos en las colecciones de prueba, genera las curvas de exhaustividad - precisión.

Al estar desarrollado en el lenguaje C y al distribuirse su código gratuitamente, permite utilizar todas las rutinas desarrolladas para implantar estos cuatro módulos anteriores, de tal forma que desde un programa externo se puede acceder a la información que almacena este S.R.I. sin necesidad de utilizarlo como medio para ello.

Seguidamente vamos a esbozar el mecanismo que aplica SMART a la hora de calcular el peso de los términos de los documentos y las consultas, ya que se reseñará posteriormente en el desarrollo de esta memoria. El proceso de generación de los pesos de cada término de un vector se compone de tres fases, que parten de la frecuencia del término en el documento o consulta (tf) [SB88]:

1. Normalización del tf de cada término del vector.
2. Modificación del peso calculado en la etapa anterior, generalmente con información proveniente de la colección completa, para así aumentar el peso de los términos menos comunes y disminuir el de los más comunes.
3. Normalización del vector completo.

El proceso completo de ponderación se nota mediante una palabra de tres caracteres. Cada uno de ellos representa el método empleado en la fase correspondiente, existiendo un total de cinco alternativas posibles para cada uno. Veamos como ejemplo la interpretación de los esquemas de ponderación utilizados a lo largo de los experimentos hechos con SMART en el desarrollo de esta memoria:

- Esquema *nnn*:
  1. $n \rightarrow$ No se hace ninguna conversión, dejando el valor del tf intacto.
  2. $n \rightarrow$ No se combina el tf con ninguna información de la colección.
  3. $n \rightarrow$ No se normaliza el vector completo.

  Quedando al final como peso en todos los términos el tf de cada uno.
- Esquema *ntn*:
  1. $n \rightarrow$ No se hace ninguna conversión, dejando el valor del tf intacto.
  2. $t \rightarrow$ Se calcula el peso $tf \cdot idf$ del término.
  3. $n \rightarrow$ No se normaliza el vector completo.

  El peso final será el $tf \cdot idf$ de cada término.
- Esquema *ntc*:
  1. $n \rightarrow$ No se hace ninguna conversión, dejando el valor del tf intacto.
  2. $t \rightarrow$ Se calcula el peso $tf \cdot idf$ del término.
  3. $c \rightarrow$ Se normaliza el vector completo por la raíz cuadrada de la suma de los $tf \cdot idf$ al cuadrado.

  El peso final será el $tf \cdot idf$ normalizado según la medida del coseno.

1.1.6. Colecciones estándar de prueba.

A la hora de medir la calidad recuperadora de un S.R.I. nos encontramos con el problema de que los documentos relevantes a una consulta son totalmente desconocidos y, por tanto, no se pueden determinar exactamente las curvas de exhaustividad y precisión. Los S.R.I. experimentales deben, de alguna forma, tener ese conocimiento para conseguir una evolución positiva en su comportamiento. Por esto se suelen utilizar unas colecciones documentales de prueba, que constan de:

- Un conjunto de documentos, que contienen información como el título, autor, fecha y resumen, incluso las citas de unos a otros documentos.

Colección	Núm. documentos	Núm. términos.	Núm. consultas.	Ámbito
ADI	82	828	35	Ciencias información
CACM	3204	7562	64	Informática
CISI	1460	4985	112	Biblioteconomía
CRANFIELD	1398	3857	225	Aeronáutica
MEDLARS	1033	7170	30	Medicina

Cuadro 1.3.: Principales características de las colecciones estándar de prueba utilizadas.

- Un conjunto de consultas, efectuadas en lenguaje natural o en alguno formal, como puede ser el booleano.
- Un conjunto de juicios de relevancia para cada consulta, es decir, el conjunto de documentos que se consideran relevantes para todas las consultas contenidas en el segundo conjunto.

A pesar de ser una práctica ampliamente utilizada la de evaluar los S.R.I. aplicándolos a colecciones de prueba, Turtle argumenta como inconvenientes [Tur90] que no poseen un tamaño acorde con el que tienen las colecciones reales, que las representaciones de los documentos se obtienen de resúmenes y no del texto completo, y que los juicios de relevancia se ven afectados por una gran cantidad de factores.

En esta memoria presentamos resultados de los experimentos realizados con cinco de estas colecciones: ADI, CACM, CISI, CRANFIELD y MEDLARS. El número de documentos, términos y consultas, así como el ámbito en el que se enmarcan, se muestran en la tabla 1.3.

Con objeto de tener un conocimiento mayor de cada una de estas colecciones, hemos realizado un estudio estadístico para así determinar sus principales propiedades. Este análisis se ha realizado partiendo para cada documento y consulta, del número de términos por el que ha sido indexado y el idf medio, máximo y mínimo. Se ha elegido el idf como peso de los términos ya que da una medida de la generalidad o especificidad del término a través de toda la colección. A partir de esa información se ha calculado, para cada medida, la media, desviación típica, error típico de la media, los valores máximo y mínimo y los percentiles 25, 50 (mediana) y 75. Los resultados de este estudio aparecen en las tablas del anexo A.

1.2. Introducción a las redes bayesianas: conceptos básicos, aprendizaje y propagación.

En esta sección vamos a introducir la herramienta básica sobre la que se fundamentará el modelo de recuperación que se presenta en esta memoria: las *redes bayesianas*. Describiremos en primer lugar qué es una red bayesiana y cómo está formada, para pasar seguidamente a esbozar los algoritmos existentes para su uso y describir algunos mecanismos para su construcción. Un estudio más amplio y detallado de las mismas, los diversos formalismos existentes, así como su aspecto axiomático pueden encontrarse en [Pea88, Nea90, Pea93, CGH96, Jen96].

Cuando el conocimiento que manejamos es incierto, las *Redes de Creencia* se presentan como una atractiva solución a este problema. Una red de creencia es una estructura gráfica (un grafo) que, de forma explícita, representa un conjunto de variables y las relaciones de dependencia e independencia entre éstas.

La topología del grafo se puede considerar como una representación cualitativa del conocimiento. Además, una red de creencia nos permite representar el conocimiento cuantitativo midiendo la fuerza de las relaciones entre las variables. Cuando el conocimiento cuantitativo viene determinado por un conjunto de distribuciones de probabilidad, a este tipo de redes se las denomina *redes bayesianas*, y serán el tipo de redes que manejaremos en esta memoria.

Son varias las ventajas que presentan las redes bayesianas como mecanismo para modelar y, posteriormente, usar el conocimiento. Podemos destacar:

1. Son un formalismo genérico, capaz de adaptarse a un gran número de aplicaciones prácticas, incluyendo tanto labores de ingeniería del conocimiento como labores de inferencia estadística.

2. Todo el conocimiento se expresa con el mismo formato, próximo a la forma que tiene el ser humano de representar el conocimiento. Hace uso de relaciones de relevancia o causalidad entre variables.

3. Permiten tener una visión global del problema que estamos resolviendo.

4. Disponen de mecanismos para realizar distintas tareas de razonamiento (inferencia, abducción, toma de decisiones, ...) de forma eficiente.

5. Las conclusiones que se obtienen son fáciles de interpretar, tienen capacidad de explicar dichas conclusiones, así como de modificarlas ante la llegada de nueva información.

Para esta introducción sobre redes bayesianas utilizaremos la notación que a continuación especificamos, aunque cuando estemos centrados en el ámbito de la recuperación de información se adaptará ligeramente. Vamos a considerar un conjunto finito $\mathcal{U}$ de variables aleatorias discretas, donde cada variable $x \in \mathcal{U}$ puede tomar valores de un dominio finito. Utilizaremos letras minúsculas o griegas (p.ej. x, y, z) para designar variables individuales y mayúsculas para notar conjuntos de variables (p.ej. X, Y, Z). Utilizaremos las correspondientes negritas para notar la asignación de un valor específico o configuración para una variable o un conjunto de variables, respectivamente $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ y $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$.

1.2.1. Composición de una red bayesiana.

Una red bayesiana nos va a permitir representar nuestro conocimiento sobre un determinado problema a través de estructuras gráficas (*Grafos Dirigidos Acíclicos, G.D.A.*), donde los nodos representan las variables y los arcos representan relaciones de causalidad, relevancia o dependencia entre ellas. Si analizamos topológicamente la red, obtenemos una representación cualitativa del conocimiento mediante un conjunto de relaciones de dependencia e independencia entre variables. Este análisis nos permite obtener una interpretación semántica de la red, esto es, para un determinado problema, podemos leer y entender las relaciones de relevancia o de causalidad entre variables. Una relación de relevancia entre dos variables, x e y, implica una modificación en la creencia sobre x, dado que se conoce el valor que toma la variable y. Análogamente, una relación de independencia entre x e y se interpreta como una no ganancia de información (no se modifica la creencia) al conocer y.

Definición 1.1 (Dependencia e independencia marginal). *Sean X e Y dos conjuntos disjuntos de variables, entonces X e Y se dicen marginalmente independientes, si y sólo si*

$$p(\mathbf{X}|\mathbf{Y}) = p(\mathbf{X})$$

$\forall \mathbf{X}, \mathbf{Y}$ *para los que $p(\mathbf{Y}) > 0$, y se nota como $I(X,Y)$. En otro caso X e Y se dicen marginalmente dependientes y lo notaremos como $\neg I(X,Y)$.*

Por otro lado, dadas tres variables x, y, z una relación de independencia condicional se puede interpretar como que una vez que se conoce z, el conocimiento de y no aporta nada sobre el conocimiento que tenemos de x.

Definición 1.2 (Dependencia e independencia condicional). *Sean X,Y y Z tres conjuntos disjuntos de variables, entonces X e Y se dicen condicionalmente independientes dado Z, si y sólo si*

$$p(\mathbf{X}|\mathbf{YZ}) = p(\mathbf{X}|\mathbf{Z})$$

$\forall \mathbf{X}, \mathbf{Y}, \mathbf{Z}$ *para los que $p(\mathbf{YZ}) > 0$, y se nota como $I(X,Y|Z)$. En otro caso X e Y se dicen condicionalmente dependientes dado Z y lo notaremos como $\neg I(X,Y|Z)$.*

La independencia marginal puede ser tratada como un caso particular de la independencia condicional. Que X e Y sean marginalmente independientes se notará mediante $I(X,Y|\emptyset)$, donde $\emptyset$ es el conjunto vacío.

Si nos centramos en redes bayesianas, el conjunto de relaciones de independencia en el modelo gráfico se pueden obtener analizando la topología del grafo. El criterio de *d-separación* puede ser utilizado para este propósito.

Definición 1.3 (d-separación). *Dado un G.D.A. G, un camino no dirigido c (una secuencia de nodos adyacentes sin tener en cuenta la dirección de los enlaces) entre los nodos x e y se dice que está bloqueado por un conjunto de nodos Z, si existe algún nodo γ en c tal que*

- *$\gamma \in Z$, donde γ es cualquier nodo de c que no tiene arcos cabeza a cabeza (dos arcos que inciden sobre el mismo nodo) o bien,*
- *$\gamma \notin Z$, ni ningún descendiente de γ está en Z, γ tiene arcos cabeza a cabeza en c.*

Un camino que no se encuentra bloqueado se dice que está activo. Dos subconjuntos de nodos, X e Y, se dice que están d-separados por Z y se nota $\langle X,Y|Z\rangle_G^d$, si todos los caminos entre los nodos de X y los de Y están bloqueados por Z.

El concepto de independencia puede utilizarse para obtener una representación eficiente de la información cualitativa. Así, cuando hablamos de redes Bayesianas, el conocimiento cuantitativo viene determinado por una distribución de probabilidad conjunta sobre el conjunto de variables consideradas, $\mathcal{U} = \{x_1,\ldots,x_n\}$. La regla de la cadena nos permite representar la distribución de probabilidad, $p(x_1,x_2,\ldots,x_n)$, como

$$p(x_1,x_2,\ldots,x_n) = p(x_n \mid x_{n-1},\ldots,x_1)\ldots p(x_3 \mid x_2,x_1)p(x_2 \mid x_1)P(x_1)$$

Como consecuencia del concepto de d-separación, una propiedad importante que se verifica en estos modelos es que, conocidas las causas directas de una variable x_i, ésta es condicionalmente independiente del resto de variables, excepto sus consecuentes. Por tanto, la relación anterior se puede expresar como:

$$p(x_1,x_2,\ldots,x_n) = p(x_n \mid \pi(x_n))\ldots p(x_3 \mid \pi(x_3))p(x_2 \mid \pi(x_2))p(x_1)$$

con $\Pi(x_i)$ representando el conjunto de causas directas de x_i, padres de x_i en el grafo. Como consecuencia, la distribución de probabilidad conjunta se puede recuperar a través de la siguiente expresión:

$$p(x_1,x_2,\ldots,x_n) = \prod_{i=1}^{n} p(x_i \mid \Pi(x_i))$$

Para almacenar la información cuantitativa sólo necesitamos conocer para cada nodo, una distribución de probabilidad condicional para cada configuración de los padres, consiguiendo un ahorro considerable en el espacio requerido.

1.2.1.1. Ejemplo de una red bayesiana.

Con objeto de ilustrar el concepto de red bayesiana, vamos a centrarnos en el siguiente ejemplo [LS88], enmarcado en el ámbito médico, en donde observaremos de forma práctica lo expuesto en esta sección.

La disnea (dificultad al respirar) es una enfermedad que puede producirse por padecer cáncer de pulmón, tuberculosis, bronquitis, ninguna de ellas o más de una. Una visita reciente a Asia incrementa la posibilidad de sufrir tuberculosis, mientras que fumar es otro factor de riesgo tanto para el cáncer de pulmón, como para la bronquitis. Los resultados de una prueba de rayos X no discriminan entre el cáncer de pulmón y la tuberculosis, ni sobre la presencia o ausencia de disnea.

De esta descripción, se seleccionan las siguientes variables aleatorias (entre paréntesis se indica la correspondiente abreviatura para facilitar la notación), determinando los valores que podrán tomar:

- *¿HA VISITADO ASIA? (V)* $\in \{v, \bar{v}\}$
- *¿TUBERCULOSIS? (T)* $\in \{t, \bar{t}\}$
- *¿CÁNCER DE PULMÓN? (C)* $\in \{c, \bar{c}\}$
- *¿FUMA? (F)* $\in \{f, \bar{f}\}$
- *¿BRONQUITIS? (B)* $\in \{b, \bar{b}\}$
- *¿TUBERCULOSIS O CÁNCER DE PULMÓN?* [2] *(TC)* $\in \{tc, \overline{tc}\}$
- *¿RAYOS X POSITIVOS? (R)* $\in \{r, \bar{r}\}$
- *¿DISNEA? (D)* $\in \{d, \bar{d}\}$

Como se puede apreciar, las variables son binarias y el primer valor corresponde a una respuesta afirmativa a la pregunta que representa la variable, y el segundo, una negativa.

El conocimiento médico expresado en el párrafo anterior puede verse representado por la red bayesiana de la figura 1.4.

Centrémonos inicialmente en el conocimiento cualitativo que está almacenado en la red. Fijémonos en el subgrafo:

$$\text{¿FUMA?} \rightarrow \text{¿CÁNCER DE PULMÓN?} \rightarrow \text{TC} \rightarrow \text{¿BRONQUITIS?}$$

Las relaciones de dependencia que extraemos son: el conocer o no si un paciente fuma determina el padecer o no cáncer de pulmón. Este hecho determinará, a su vez, el conocimiento sobre la variable *¿TUBERCULOSIS O CÁNCER DE PULMÓN?* y, de manera sucesiva, influirá en el resultado de la prueba de rayos X. Por tanto, podemos decir que las variables *¿FUMA?* y *¿RAYOS X POSITIVOS?* son variables dependientes. En el momento en que se conoce el valor de la variable *¿CÁNCER DE PULMÓN?*, se hacen independientes: dado que se sabe que el paciente sufre cáncer de pulmón el hecho de saber si fuma o no, no nos aporta ninguna información sobre el resultado de la prueba de rayos X.

$$\text{¿CÁNCER DE PULMÓN?} \leftarrow \text{¿FUMA?} \rightarrow \text{¿BRONQUITIS?}$$

Se hace un razonamiento análogo: si sabemos que el paciente padece cáncer de pulmón, podemos suponer que tiene también bronquitis, pero en el momento en el que conocemos que

[2] Es una variable ficticia que representa el hecho de que se padece o no alguna, ambas o ninguna de las enfermedades.

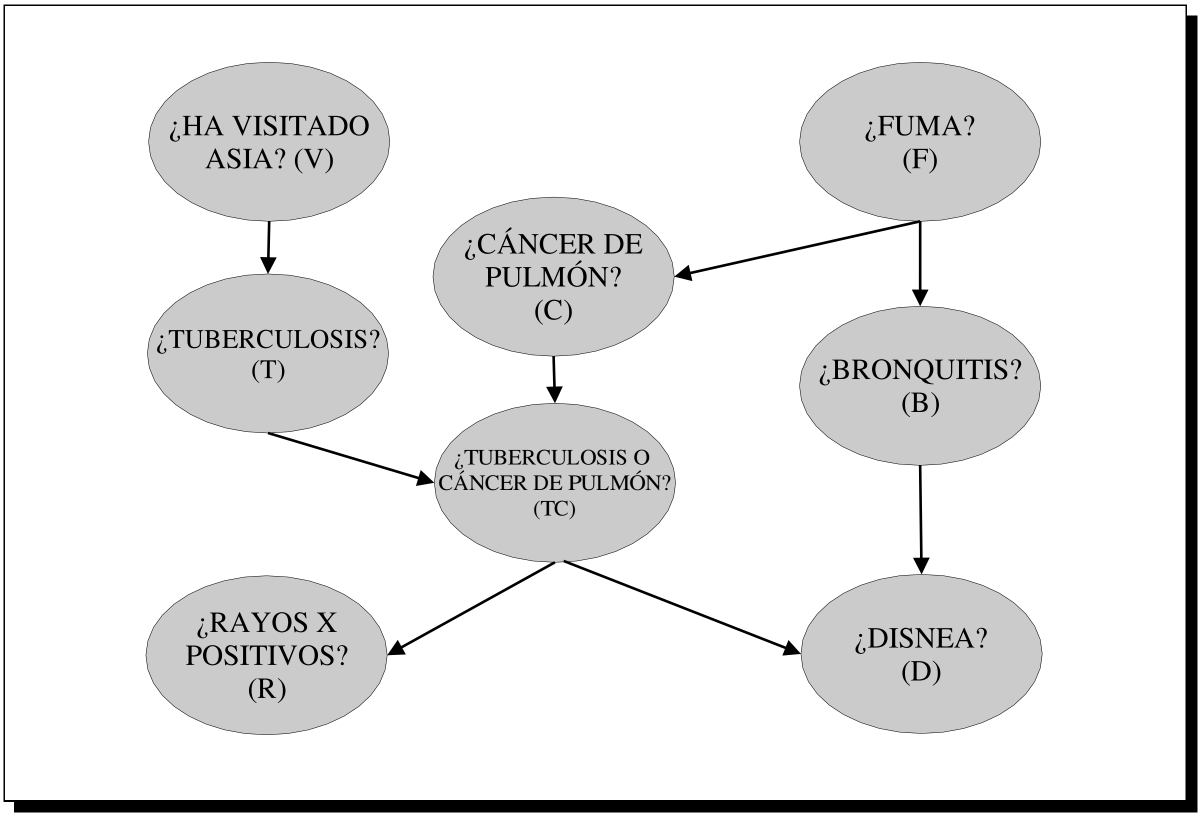


Figura 1.4.: Red bayesiana que representa el conocimiento sobre cómo padecer disnea.

fuma, el hecho de conocer que sufre cáncer no aporta información sobre que padezca o no de bronquitis. De esta manera, *¿CÁNCER DE PULMÓN?* y *¿DISNEA?* son dependientes hasta que se conoce el valor de *¿FUMA?*, momento en el cual se vuelven independientes.

El último patrón que podemos apreciar en el grafo es:

$$\text{¿TUBERCULOSIS?} \rightarrow TC \rightarrow \text{¿DISNEA?} \leftarrow \text{¿BRONQUITIS?}$$

En esta situación, el hecho de conocer que se ha contraído la tuberculosis no nos aporta información sobre el padecer o no bronquitis. Pero, en el momento en el que conocemos que el paciente sufre disnea y que padece de tuberculosis, la creencia sobre tener bronquitis aumenta. Por tanto, las variables *¿TUBERCULOSIS?* y *¿BRONQUITIS?* son independientes, pero conocido el valor de *¿DISNEA?* se hacen dependientes.

Estudiemos seguidamente el aspecto cuantitativo de la red, es decir, el que mide la fuerza de las relaciones existentes. Para cada nodo que la compone se tiene que estimar el conjunto de distribuciones de probabilidad para cada combinación de valores que toman las variables padre. En el ejemplo, podrían ser las indicadas en la tabla 1.4 (los casos en los que las variables condicionadas toman los valores negativos, se obtienen por dualidad. Así, por ejemplo, $p(\bar{d} \mid tc, b) = 1 - p(d \mid tc, b) = 0{,}1$.

Variable	Distribuciones
V	$p(v) = 0{,}01$
T	$p(t \mid v) = 0{,}05$
	$p(t \mid \bar{v}) = 0{,}01$
C	$p(c \mid f) = 0{,}10$
	$p(c \mid \bar{f}) = 0{,}01$
F	$p(f) = 0{,}5$
B	$p(b \mid f) = 0{,}6$
	$p(b \mid \bar{f}) = 0{,}3$
TC	$p(tc \mid t,c) = 1{,}0$
	$p(tc \mid t,\bar{c}) = 1{,}0$
	$p(tc \mid \bar{t},c) = 1{,}0$
	$p(tc \mid \bar{t},\bar{c}) = 0{,}0$
R	$p(r \mid tc) = 0{,}98$
	$p(r \mid \overline{tc}) = {,}05$
D	$p(d \mid tc,b) = 0{,}90$
	$p(d \mid tc,\bar{b}) = 0{,}70$
	$p(d \mid \overline{tc},b) = 0{,}80$
	$p(d \mid \overline{tc},\bar{b}) = 0{,}10$

Cuadro 1.4.: Distribuciones de probabilidad de la red ejemplo.

Como dijimos anteriormente, la distribución de probabilidad conjunta se puede generar a partir del producto de las almacenadas en la red. En nuestro ejemplo,

$$p(V,T,C,F,B,TC,R,D) =$$

$$= p(V)p(T \mid V)p(C \mid F)p(F)p(B \mid F)p(TC \mid T,C)p(R \mid TC)p(D \mid TC,B)$$

En este caso, sólo necesitamos almacenar cuarenta valores de probabilidad en lugar de los doscientos cincuenta y seis que se necesitarían para almacenar la distribución de probabilidad conjunta.

1.2.2. Tipos de redes bayesianas.

Atendiendo a la complejidad del G.D.A. subyacente a la red bayesiana, los más sencillos son las *redes simplemente conectadas*: aquéllas en las no existe más de un camino (no dirigido) que conecte dos nodos cualesquiera (grafos que no tienen ningún tipo de ciclo). Los *árboles* y los *poliárboles* son los representantes de este grupo. Los primeros tienen como característica principal que los nodos contenidos en ellos, como máximo, sólo tienen un padre, mientras que los segundos no tienen restricción en cuanto al número de éstos.

Otro tipo especial de redes bayesianas son los *grafos simples*, es decir, grafos donde cada par de nodos con un hijo en común no tienen antecesores comunes, ni uno es antecesor del otro. Cuando existen varias sucesiones de nodos (caminos) que parten de un mismo nodo origen y llegan a un mismo nodo de destino, entonces tendremos un grafo *múltiplemente conectado*.

En la figura 1.5 mostramos las topologías de estos grafos que acabamos de introducir aquí. El tipo de red bayesiana es importante para el proceso de inferencia, como veremos a continuación.

1.2.3. Inferencia en redes bayesianas.

Una vez que la base de conocimiento está creada, la utilidad de ésta es la de obtener conclusiones a la luz de la nueva información que llega. Esta nueva información se conoce con el nombre de *evidencia*, y al mecanismo de inferencia se le denomina *propagación de evidencias*, y consiste básicamente en actualizar las probabilidades de las variables representadas en el grafo teniendo en cuenta dichas evidencias.

Así, dado un conjunto de variables $\mathcal{U} = \{x_1, \cdots, x_n\}$, cuando no existe ninguna evidencia, el resultado de la propagación simplemente es la probabilidad a priori, $p(x_i)$ de que cada variable tome sus correspondientes valores. Por otro lado, cuando se dispone de evidencias, es decir, un conjunto de variables, $E \subseteq \mathcal{U}$, de las cuales se conoce el valor que toma cada una, el proceso de propagación genera como salida la probabilidad de que cada variable que no está en el conjunto de las evidencias tome cada uno de sus valores, dado que las evidencias toman unos valores concretos conocidos, probabilidades denominadas *a posteriori*: $p(\mathbf{x_i} \mid \mathbf{E})$.

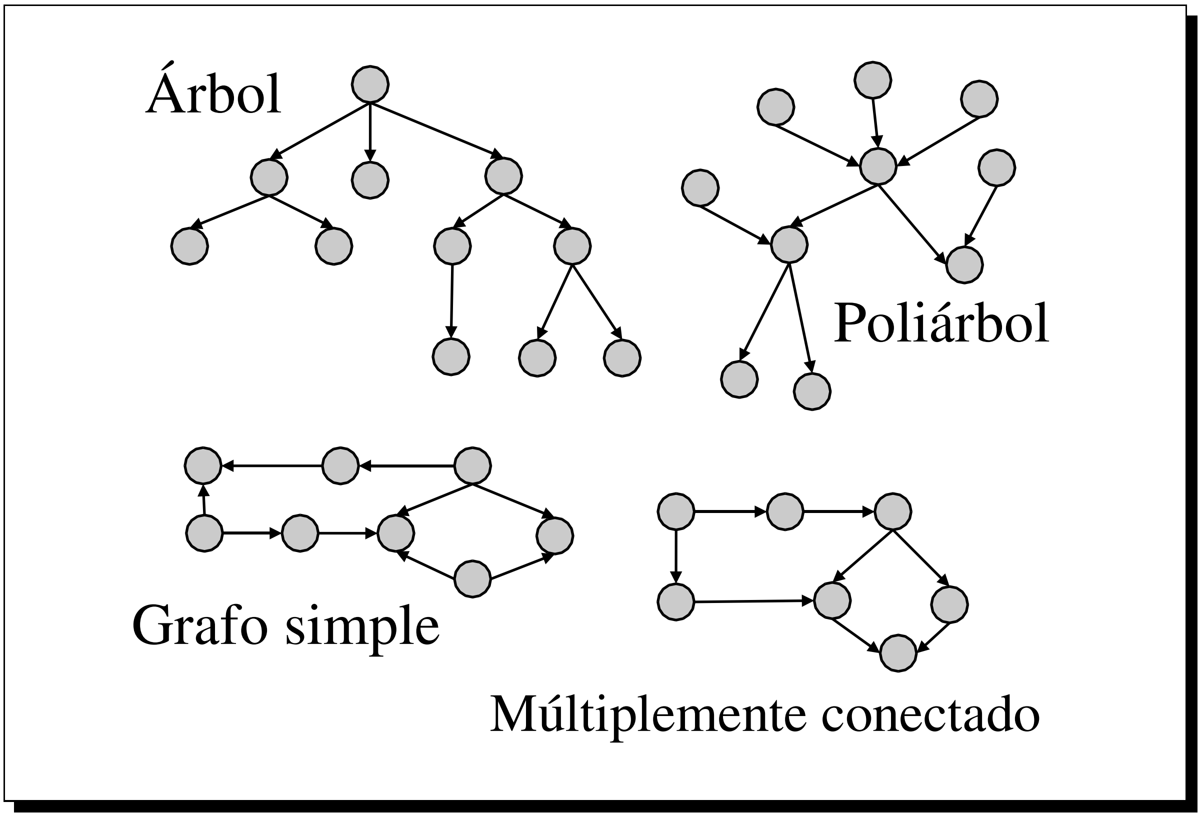


Figura 1.5.: Diferentes tipos de redes bayesianas.

En el campo de las redes bayesianas, existen tres grupos de algoritmos de propagación [CGH96]:

- Exactos: calculan las probabilidades mediante expresiones matemáticas exactas, cometiendo únicamente el error originado por el redondeo del ordenador.
- Aproximados: utilizan diferentes métodos de simulación para obtener las probabilidades.
- Simbólicos: obtienen las probabilidades en función de parámetros, que luego podrán sustituirse por valores reales.

La propagación exacta tiene el inconveniente de que es un problema NP-duro [Coo90] y, por tanto, de poca aplicación práctica en ciertos casos. Esta es la razón por la que surgieron técnicas que, a costa de perder precisión en los cálculos, obtuvieran resultados en un tiempo bastante menor. Los algoritmos aproximados se basan en la generación de una muestra aleatoria a partir de la distribución de probabilidad conjunta, para luego calcular, a partir de ella, las probabilidades de ciertos sucesos dadas las evidencias. En [CGH96, Sal98] se hace una revisión de las principales técnicas de propagación aproximada existentes. Entre estas destacamos, debido a que es la que hemos usado en algunos de nuestros experimentos, la conocida como *muestreo* *por importancia*, la cual se ejecuta en dos fases: una primera, que lleva a cabo una eliminación de variables para encontrar una aproximación de las funciones de muestreo, y una segunda, de simulación, que genera las configuraciones de las variables a partir de las funciones obtenidas en la fase anterior [SCM00, Sal98].

Volviendo a la propagación exacta, son varios los tipos de algoritmos desarrollados para realizar esta tarea de inferencia [Pea88, CGH96] atendiendo al tipo de G.D.A. donde se van a utilizar:

- Redes simplemente conectadas:
    - Método de paso de mensajes en poliárboles [Pea88]. Este método se basa en la combinación por parte de cada nodo del grafo (cálculos locales) de la información que recibe de sus padres y de sus hijos mediante un conjunto de mensajes. A su vez, dicho nodo manda mensajes a sus antecesores y descendientes directos. La propagación es un flujo de mensajes entre los diferentes nodos de la red hasta que todos los nodos han calculado de manera local la información requerida (la probabilidad a posteriori).
- Redes múltiplemente conectadas:
    - Métodos de condicionamiento [Pea86, SC91]. La idea principal es cortar esos caminos entre los nodos mediante la asignación de valores a un conjunto reducido de variables contenidas en los ciclos (se instancian esas variables). De esta forma se obtiene una red simplemente conectada sobre la que se puede aplicar el algoritmo de propagación de paso de mensajes en poliárboles.
    - Métodos de agrupamiento [LS88, JLO90, SAS94]. Construyen representaciones más simples del grafo uniendo conjuntos de nodos del grafo original, creando así los denominados *conglomerados*. Una vez que se ha formado un árbol de conglomerados se aplica el método de propagación exacta en poliárboles.
    - Métodos mixtos (condicionamiento y agrupamiento) [SC91b, SCH91]. Métodos que combinan las ventajas de ambos tipos de técnicas debido a que ninguno de los dos destaca, en cuanto a ventajas, sobre el otro.
    - Métodos orientados a un objetivo [GVP90, GVP90b]. Estos métodos se aplican cuando el número de variables sobre las que se tiene interés para calcular su probabilidad a posteriori dadas las evidencias es reducido. En este caso, algunas variables del grafo pueden ser irrelevantes para este objetivo, por lo que pueden ser eliminadas, consiguiendo un modelo equivalente más sencillo que sólo contenga nodos relevantes y donde será más fácil propagar.

Debido a que hemos optado por el poliárbol como base para varios de los modelos de recuperación que presentamos en esta memoria, hemos elegido como método de propagación exacta la basada en paso de mensajes. Seguidamente vamos a exponer con algo más de profundidad este método siguiendo la explicación que del mismo hacen Castillo y col. en [CGH96](también aparece detalladamente explicado en [Pea86, Pea88, Her98]).

Algoritmo de propagación exacta en poliárboles mediante paso de mensajes de Pearl
Este algoritmo de propagación tiene como característica principal que posee una complejidad proporcional al tamaño de la red (número de nodos y arcos) y hace uso de la propiedad de los poliárboles que indica que cualquier nodo de éste lo divide en dos poliárboles inconexos: el formado por sus ascendientes (los nodos que estén por encima de él) y el compuesto por sus descendientes (los que estén por debajo), como se puede ver en la figura 1.6 con el nodo x_4.

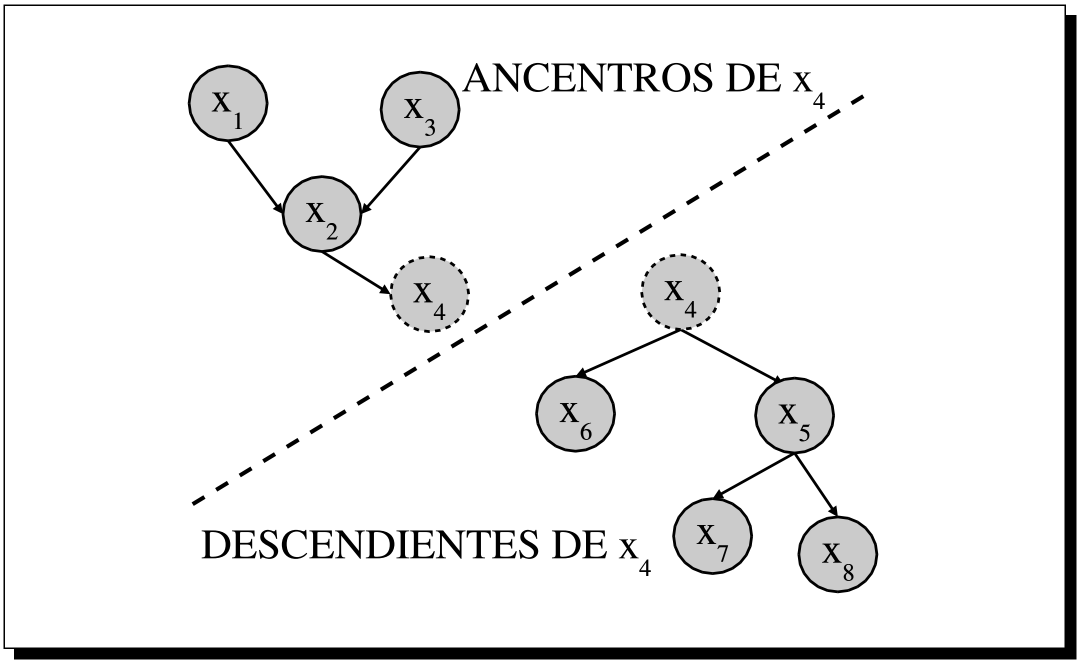


Figura 1.6.: Los dos poliárboles (el de ancestros y el de descendientes) que surgen a partir de un nodo.

Así, teniendo en cuenta dicha propiedad, el proceso de propagación (dada una configuración $\mathbf{E}$ para un conjunto de variables evidencias E, calcula $p(\mathbf{x_i} \mid \mathbf{E}), \forall x_i \in G$) se basa en la combinación de la información que llega de ambos poliárboles, mediante el envío de mensajes de un lugar a otro. El conjunto de evidencias E, para un nodo cualquiera x_i se descompone en dos subconjuntos: E_i^+, que representa al subconjunto de evidencias que están situadas en el poliárbol de ancestros de x_i, y E_i^-, el subconjunto de evidencias localizadas, en este caso, en el poliárbol de descendientes de x_i. En la figura 1.7 vemos cada uno de estos conjuntos para el grafo de la figura 1.6.

La probabilidad a posteriori de x_i se obtiene aplicando la siguiente expresión:

$$p(\mathbf{x_i} \mid \mathbf{E}) = \frac{\lambda_i(\mathbf{x_i}) \cdot \rho_\mathbf{i}(\mathbf{x_i})}{k}, \tag{1.14}$$

donde, $\lambda_i(\mathbf{x_i})$ es la información procedente de los hijos de x_i, es decir, $p(\mathbf{E_i^-} \mid x_i)$, y $\rho_i(\mathbf{x_i})$ la información procedente de los padres, o lo que es lo mismo, $p(x_i \mid \mathbf{E_i^+})$, y por último, k es una

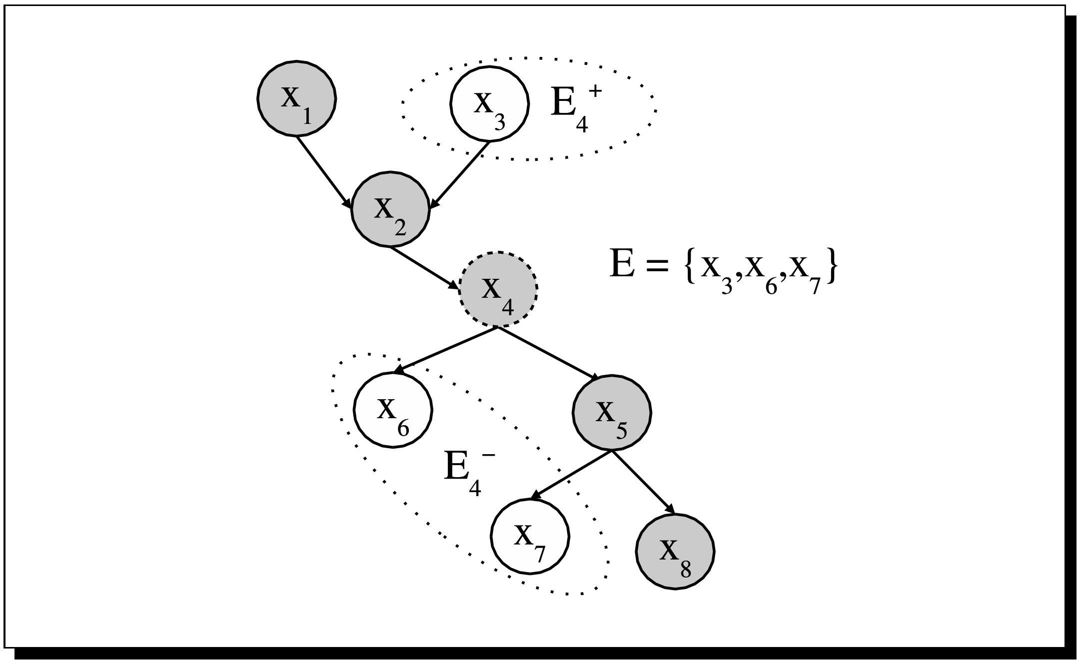


Figura 1.7.: Conjuntos de evidencias de los ancestros y descendientes.

constante de normalización. Por tanto, para calcular la probabilidad a posteriori de un valor concreto de una variable se combinan dichas informaciones mediante el producto de ambas.

El siguiente problema a plantear será cómo cada nodo forma los mensajes λ y ρ: por un lado, el nodo espera a que todos los mensajes λ que recibe directamente de sus hijos estén calculados y, posteriormente, los combina, y por otro lado, realizando la misma operación con los mensajes ρ que le llegan a partir de cada uno de sus padres. A su vez, ese mismo nodo calculará y mandará mensajes λ a sus padres y ρ a sus hijos. Veámoslo más detalladamente: dado un nodo cualquiera x_i, éste tiene p nodos padre, representados por $\Pi(x_i) = \{u_1, \ldots, u_p\}$ y h nodos hijo, $H = \{y_1, \ldots, y_h\}$. El conjunto de evidencias E_i^+ se puede descomponer en p subconjuntos disjuntos, uno para cada padre:

$$E_i^+ = \{E_{u_1,x_i}^+, \ldots, E_{u_p,x_i}^+\},$$

donde la evidencia E_{u_j,x_i}^+ es el subconjunto de E_i^+ contenido en el subgrafo asociado al nodo u_j cuando se elimina la arista $u_j \rightarrow x_i$. De manera análoga, E_i^-, se puede dividir en h subconjuntos distintos asociados al nodo x_i:

$$E_i^+ = \{E_{x_i,y_1}^+, \ldots, E_{x_i,y_h}^-\},$$

siendo E_{x_i,y_j}^+ el subconjunto de E_i^- contenido en el subgrafo asociado al nodo y_j cuando se elimina la arista $x_i \rightarrow y_j$. En la figura 1.8 se observa la situación que acabamos de describir.

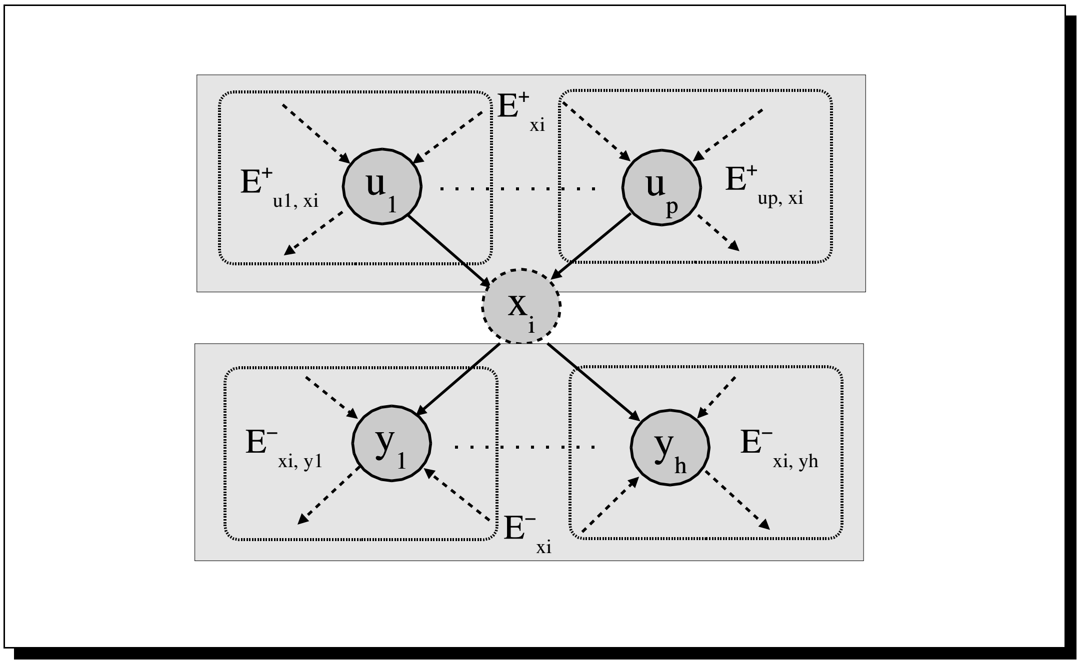


Figura 1.8.: Subconjuntos de evidencias asociados a los padres y a los hijos de x_i.

Sea $\pi(x_i)$ una configuración de valores para los padres de x_i. El cálculo de ρ se hace de la siguiente forma:

$$\rho_i(\mathbf{x_i}) = \sum_{\pi(x_i)} p(\mathbf{x_i} \mid \pi(x_i) \cup \mathbf{E_i^+}) \prod_{j=1}^{p} \rho_{u_j x_i}(\mathbf{u_j}), \tag{1.15}$$

$\rho_{u_j x_i}(\mathbf{u_j})$ es el mensaje ρ que el nodo u_j envía a su hijo x_i y representa $p(\mathbf{u_j} \cup \mathbf{E^+_{u_j,x_i}})$.

La función λ tiene la siguiente expresión:

$$\lambda_i(\mathbf{x_i}) = \prod_{j=1}^{h} \lambda_{y_j,x_i}(\mathbf{x_i}), \tag{1.16}$$

donde $\lambda_{y_j,x_i}(\mathbf{x_i}) = p(\mathbf{E^-_{x_i,y_j}} \mid \mathbf{x_i})$

Por último, una vez que el nodo x_i ha calculado $\lambda_i(\mathbf{x_i})$ y $\rho_i(\mathbf{x_i})$, debe calcular los mensajes que les mandará a sus padres y a sus hijos. Por un lado, el mensaje que le manda a cada hijo y_j es el siguiente:

$$\rho_{x_i,y_j}(\mathbf{x_i}) \propto \rho_i(\mathbf{x_i}) \prod_{k \neq j} \lambda_{y_k,x_i}(\mathbf{x_i}) \tag{1.17}$$

Por otro lado, el cálculo del mensaje que x_i manda a sus padres, $\lambda_{x_i,u_j}(\mathbf{u_j})$, se hace de acuerdo a:

$$\lambda_{x_i,u_j}(\mathbf{u_j}) = \sum_{\mathbf{x_i}} \lambda_{x_i}(\mathbf{x_i}) \sum_{\pi(x_i)-\mathbf{u_j}} p(\mathbf{x_i} \mid \pi(x_i)) \prod_{k=1,\ldots,p/k\neq j} \rho_{u_k,x_i}(\mathbf{u_k}) \tag{1.18}$$

Todos los mensajes que un nodo del grafo recibe y manda podemos verlos en la figura 1.9.

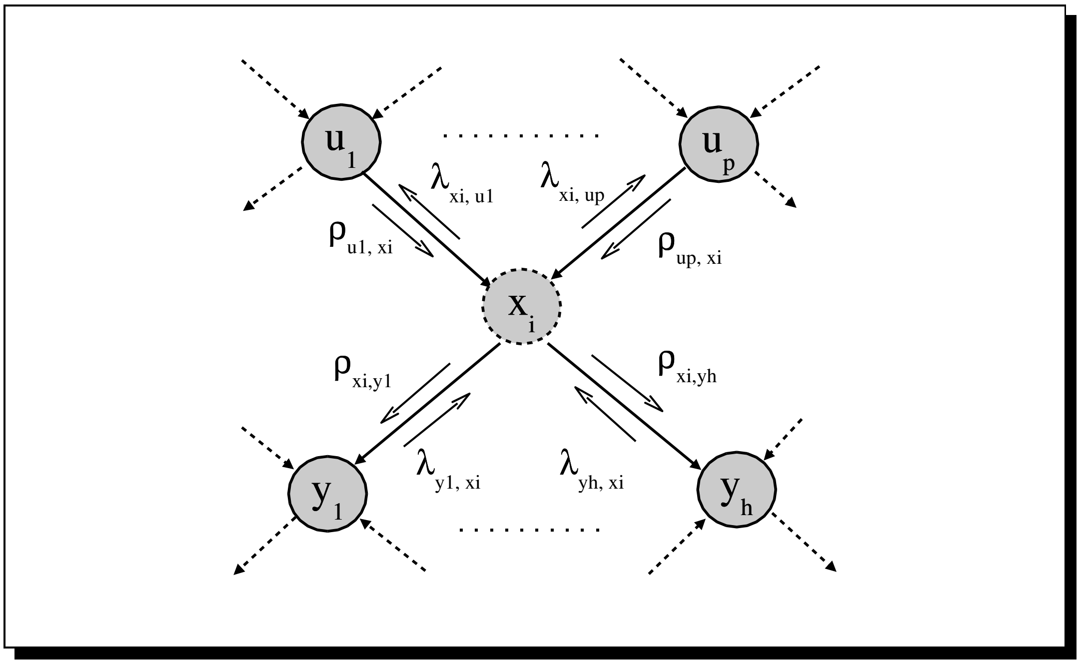


Figura 1.9.: Mensajes λ y ρ que manda y recibe un nodo x_i.

En realidad, los mensajes λ y ρ de cada nodo son vectores de tantos elementos como número de casos tengan las variables correspondientes. El valor inicial de ambos tipos de mensajes cuando el nodo es una variable evidencia será un uno en la posición correspondiente al valor que toma dicha variable evidencia, y un cero en el resto. Para aquellos nodos que no tengan padres, es decir, nodos raíces, el vector ρ estará compuesto por las probabilidades a priori de cada uno de los casos que puede tomar la variable correspondiente. Y por último, para los nodos hoja, es decir, los que no tienen hijos, cada una de las posiciones del vector λ se inician a uno.

El proceso general, para cada nodo x_i del grafo, una vez se han asignado los valores iniciales a los mensajes es el siguiente:

1. Recibir todos los mensajes ρ de los padres y calcular $\rho_i(x_i)$ según la ecuación (1.15).
2. Recibir todos los mensajes λ de los hijos y calcular $\lambda_i(x_i)$ según la ecuación (1.16).
3. Calcular y enviar, para cada hijo y_j del nodo x_i el mensaje $\rho_{x_i,y_j}(x_i)$ por medio de la expresión (1.17).
4. Calcular y enviar, para cada padre u_j del nodo x_i el mensaje $\lambda_{x_i,u_j}(u_i)$ utilizando la expresión (1.18).

5. Repetir estos pasos hasta que todos los nodos distintos a las evidencias tengan calculados sus mensajes ρ y λ.

6. Para cada nodo, calcular la probabilidad de x_i dadas las evidencias $\mathbf{E}$ según la ecuación (1.14).

Existe un cuarto tipo de nodo especial, conocido como *nodo imaginario* (en inglés y siguiendo la terminología de Pearl *dummy node*), el cual se asocia a un nodo x_i y representa una evidencia virtual sobre él, es decir, modela la situación en que no hay una evidencia clara sobre x_i. Este nodo imaginario se sitúa en el grafo como hijo del nodo x_i y mandará a su padre un mensaje λ compuesto por verosimilitudes.

Veamos un ejemplo. Supongamos que una variable x_i cualquiera puede tomar los valores $\{v_1, v_2\}$. Tenemos una creencia diez veces mayor en $x_i = v_1$ que en $x_i = v_2$, lo que se puede representar como sigue:

$$p(\text{Observación} \mid v_2) : p(\text{Observación} \mid v_1) = 1 : 10$$

donde $p(\text{Observación} \mid b)$ representa la probabilidad de obtener una observación dado que conocemos que x_i toma el valor v_2 (análogamente para $p(\text{Observación} \mid v_1)$).

Así, el nodo imaginario creado como hijo de x_i le mandará un mensaje λ con dos verosimilitudes:

$$\lambda(x_i) = a \cdot (p(\text{Observación} \mid v_2), p(\text{Observación} \mid v_1))$$

siendo a cualquier constante que nos servirá para normalizar los valores del vector (generalmente se normaliza el vector dividiendo por el mayor de los valores que contiene):

$$\lambda(x_i) = a \cdot (\frac{p(\text{Observación} \mid v_2)}{p(\text{Observación} \mid v_1)}, 1)$$

La existencia de este tipo de nodos será la base para la técnica de la *instanciación parcial*, que presentaremos en el capítulo 3 y para el método de realimentación de relevancia, desarrollado específicamente para el modelo que expondremos en el capítulo 3.

1.2.4. Construcción de redes bayesianas.

El siguiente problema que se plantea es estudiar cómo se construye una red de creencia. Una posibilidad es que el ingeniero del conocimiento construya la red con la ayuda de expertos humanos en el problema. Sin embargo, cuando el experto tiene un conocimiento parcial sobre el problema, esta aproximación es problemática. En cualquier caso, construir este tipo de redes con la ayuda de expertos humanos es una tarea que requiere una gran cantidad de tiempo y esfuerzo. Por ello, es deseable tener técnicas automáticas que nos permitan agilizar este proceso. Este tipo de técnicas se basan en utilizar la información que se obtiene a partir de una base de datos. Además, cada vez es más usual poder encontrar grandes bases de datos disponibles, por lo que los algoritmos de aprendizaje automático representan una herramienta útil en la fase de construcción de este tipo de estructuras, a pesar de ser una tarea NP-dura [CGH94].

En esta sección trataremos las técnicas cuyo objetivo es el de recuperar la red que es capaz de reproducir un conjunto de datos. En general, estas técnicas asumen que la base de datos es una representación de la distribución de probabilidad que sigue la población, en lugar de una muestra de la misma, y su objetivo es el de encontrar la red bayesiana que mejor represente el conjunto de datos.

Estas técnicas pueden clasificarse en dos grandes grupos [Cam98b]:

- Las basadas en detección de independencias, que estudian las relaciones de independencia existentes a partir de los datos disponibles y tratan de encontrar una red que represente dichas relaciones.
- Las basadas en funciones de evaluación y técnicas de búsqueda heurística, cuyo objetivo es encontrar un grafo que represente lo más fielmente posible los datos, pero con el menor número de arcos posible.

Existen también algoritmos de aprendizaje híbridos que utilizan de forma conjunta ambas técnicas, como es el caso del algoritmo *BENEDICT* [AC96, AC00, AC01] para grafos generales, que utiliza una métrica específica y un método de búsqueda, pero también emplea las relaciones de independencia representadas en la red para definir la métrica, y utiliza tests de independencia para limitar el proceso de búsqueda.

Veamos con algo más de detalle los dos primeros grupos.

Algoritmos de aprendizaje basados en detección de Independencias. Estos algoritmos pueden tener como entrada el conjunto de relaciones de independencia existentes entre las variables del problema (a través de los datos con los que se cuenta), una distribución de probabilidad sobre la que se comprueban dichas relaciones o una base de datos con la que se estima la veracidad o no de las relaciones de independencia mediante tests estadísticos de independencia condicional.

Las diferencias entre los algoritmos de este tipo se establecen normalmente en cuanto al costo de los tests de independencias, a la fiabilidad de dichos tests y al tipo de grafo que recuperan, entre otras. Con respecto a este último criterio, existen muy diversos algoritmos de aprendizaje de este grupo que generan diferentes topologías:

- Árboles [GPP90, Cam98].
- Poliárboles [HC93b, Cam98].
- Otros grafos simples [GPP93, CH97].
- Grafos generales [SGS91, SGS93, VP90, CH00].

Algoritmos de aprendizaje basados en funciones de evaluación y heurísticas. Estos algoritmos recorren un espacio de búsqueda de grafos y determinan la calidad del grafo correspondiente usando una métrica. Poseen implícita o explícitamente los siguientes tres elementos:

1. Una medida de calidad que nos permita seleccionar la mejor estructura entre un conjunto de ellas.
2. Una heurística de búsqueda para seleccionar, de entre el conjunto de posibles estructuras por comparar, una de ellas. Generalmente, esta heurística es de tipo "ávido" (greedy),
3. Un método para obtener la información cuantitativa (distribuciones de probabilidad) de la estructura resultante.

Centrándonos en la medidas de calidad o métricas, podemos agruparlas según estén basadas en:

- Métodos bayesianos: tratan de maximizar la probabilidad de obtener una determinada estructura condicionada a la base de datos de que se dispone, utilizando para ello la fórmula de Bayes. Como ejemplo más destacado de un algoritmo de este tipo podemos citar el *Algoritmo K2* [CH92]: utiliza como métrica (debido a que se aseguran ciertas condiciones) una fórmula que establece la probabilidad conjunta de un grafo G y una base de datos BD. Mediante una búsqueda local va modificando el grafo, inicialmente vacío, de tal forma que se vaya incrementando la probabilidad de la estructura resultante.
- Descripción de longitud mínima: la mejor representación de un conjunto de datos es aquella que minimiza las longitudes de codificación del modelo y de los datos dado el modelo. En este caso, los modelos más complejos (los que representan más fielmente los datos) son las redes densamente conectadas (todos los nodos están conectados con todos), pero ofrecen dificultades computacionales y de compresión. Así, el objetivo de estos algoritmos es encontrar redes menos precisas pero más simples utilizando como métrica el principio de descripción de longitud mínima.
- Entropía: tratan de encontrar la red cuya entropía cruzada con los datos sea mínima. La entropía se puede considerar como una forma de medir el grado de dependencia entre variables, y en este sentido estos métodos buscan configuraciones que favorezcan la presencia de conexiones entre variables que manifiesten un alto grado de dependencia.

Como ejemplo de algoritmos de aprendizaje basados en métricas podemos citar [CL68, HC90, CH91, CH92, HGC95, Cam98, CP01, CP01b]

Algoritmos de construcción de poliárboles.

Seguidamente vamos a estudiar algo más profundamente los algoritmos de aprendizaje que basan su medida en la entropía y construyen redes simplemente conectadas, por ser los que fundamentalmente hemos utilizado en el desarrollo de esta memoria.

Como medida de calidad, estos algoritmos utilizan, entre otras, una medida de la distancia entre la distribución de probabilidad obtenida de los datos, P (la consideran la distribución real), y la distribución que se obtiene al considerar una estructura simplemente conectada P^T, como el producto de n distribuciones de probabilidad condicionadas. El objetivo que persiguen es el de encontrar aquella distribución P^T que mejor se adecue a la distribución real P. Para ello, utiliza como criterio de bondad en el ajuste una medida distancia entre las dos distribuciones P^T y P, la medida de Entropía de Kullback-Leibler [KL51]:

$$D(P,P^T) = \sum_{\mathbf{x_1},\ldots,\mathbf{x_n}} P(\mathbf{x_1},\ldots,\mathbf{x_n}) \log \frac{P(\mathbf{x_1},\ldots,\mathbf{x_n})}{P^T(\mathbf{x_1},\ldots,\mathbf{x_n})} \tag{1.19}$$

con $\mathbf{x_1},\ldots,\mathbf{x_n}$ representando todos los posibles casos de las variables $x_1,\ldots,x_n$. El algoritmo de búsqueda trata de minimizar la distancia $D(P,P^T)$. Para ello, es suficiente con proyectar P en un árbol generador de costo máximo, con lo que en este caso el proceso de búsqueda se realiza de forma implícita. Para cada arista (x_i,x_j) se define el costo como la medida de información mutua entre las variables, esto es, la *medida de la información mutua esperada*, [KL51], $I(x_j,x_j)$, calculada mediante la ecuación:

$$I(x_i,x_j) = \sum_{\mathbf{x_i},\mathbf{x_j}} P(\mathbf{x_i},\mathbf{x_j}) \log \frac{P(\mathbf{x_i},\mathbf{x_j})}{P(\mathbf{x_i})P(\mathbf{x_j})} \tag{1.20}$$

Entre las propiedades de la medida $I(x_i,x_j)$ cabría destacar que siempre es positiva o nula, alcanzando el mínimo (cero) cuando las dos variables son independientes. Cuanto mayor sea el valor de la cantidad de información, la dependencia entre las variables será mayor.

Veremos en primer lugar el algoritmo dado por Chow y Liu [CL68] para recuperar árboles, para posteriormente considerar dos algoritmos que nos permiten recuperar poliárboles: los propuestos por Rebane y Pearl [RP89, Pea88] y Campos [Cam98]. La razón fundamental es que un poliárbol permite representar modelos de dependencias más ricos que las estructuras arbóreas.

Para alcanzar tal objetivo, Chow y Liu proponen el conjunto de pasos que se muestra en el algoritmo 1.1.

Algoritmo 1.1 Aprendizaje de un árbol (Método de Chow y Liu).

```
1: A partir de la distribución de probabilidad conjunta observada P(x_1,...,x_n) calcular, para
   cada par de variables (x_i,x_j), la distribución marginal bidimensional P(x_i,x_j).
2: Utilizando el conjunto de pares, calcular todos los n(n−1)/2 pesos de las aristas utilizando
   la ecuación (1.20) y ordenarlos por magnitud.
3: Seleccionar el par de mayor peso y añadir una arista entre los dos nodos.
4: repite
5:    Seleccionar la siguiente arista de mayor peso y añadirla al grafo, salvo que forme un
      ciclo, en cuyo caso se elimina y se toma el siguiente par de mayor peso.
6: hasta que n−1 aristas hayan sido incluidas.
```

Este algoritmo puede generar, dada una determinada distribución de probabilidad P, distintos árboles dependiendo del orden con el que se seleccionen los arcos de igual peso. Las ventajas que presenta este algoritmo son las siguientes: para calcular la cantidad de información (ecuación (1.20)) sólo se utilizan distribuciones conjuntas bidimensionales, las cuales pueden ser calculadas de forma eficiente y fiable a partir de un número no demasiado elevado de datos. Además, el algoritmo se ejecuta en un orden $O(n^2 \log n)$, utilizando únicamente una comparación de pesos. Finalmente, si la distribución es representable por (isomorfa) un árbol, el algoritmo recupera el árbol que la representa.

El algoritmo de Rebane y Pearl [RP89], se puede considerar como una generalización del método de Chow y Liu. En una primera fase, el algoritmo obtiene un grafo no dirigido (utilizando el algoritmo de Chow y Liu), para posteriormente orientar el mayor número posible de aristas. La fase de orientación se basa en la siguiente propiedad: *En una estructura de poliárbol, dos nodos con un descendiente directo común son marginalmente independientes.* Por tanto, es posible distinguir, dado el subgrafo $x - y - z$, la estructura $x \rightarrow y \leftarrow z$ de las estructuras $x \leftarrow y \rightarrow z; x \rightarrow y \rightarrow z; x \leftarrow y \leftarrow z$, las cuales son probabilísticamente indistinguibles. Para ello, dada la terna $x - y - z$, podemos determinar si x y z son padres de y en base a tests de independencia marginal entre x y z. El conjunto de pasos para lograr este objetivo se muestra en el algoritmo 1.2.

Algoritmo 1.2 Aprendizaje de un poliárbol (Método de Rebane y Pearl).

Generar el árbol generador de costo máximo utilizando el algoritmo de Chow y Liu (Algoritmo 1.1).
repite
    Buscar una terna de nodos $x - z - y$ donde x e y sean marginalmente independientes. En este caso orientar x, y como padres del nodo z.
    Cuando una estructura de múltiples padres ha sido encontrada, determinar la dirección de todos sus arcos utilizando el test de independencia marginal entre sus adyacentes.
    para cada nodo que tenga al menos un arco de entrada **hacer**
        estudiar la direccionalidad del resto de los adyacentes mediante test de independencia marginal.
    fin para
hasta que no se puedan descubrir nuevas orientaciones.
si existen arcos sin orientar **entonces**
    etiquetarlos como 'indeterminados'.
fin si

Cuando la distribución $P(x_1, \ldots, x_n)$ puede ser representada mediante un poliárbol, el algoritmo recupera el esqueleto y además direcciona el mayor número de arcos posibles, detectando cuándo una variable tiene más de un padre. En cualquier otro caso, no existen garantías de que el poliárbol obtenido sea la mejor aproximación de $P(x_1, \ldots, x_n)$.

Por último, el tercer algoritmo es el diseñado por Campos [Cam98] y denominado *PA*. La principal diferencia con respecto al de Chow y Liu es la forma de calcular el peso de las aristas, ya que en éste la dependencia entre dos nodos se calcula directamente, mientras que en el PA se hace a partir de la combinación de la dependencia marginal de esos dos nodos y las dependencias condicionadas con a cada una de las variables restantes.

Por otro lado, y en cuanto a las diferencias y similitudes del algoritmo PA con respecto al de Rebane y Pearl, ambos basan la construcción de la estructura en el método de Chow y Liu. Por otro lado, Rebane y Pearl hacen la orientación como se indica en el algoritmo 1.2, mediante tests de independencia marginal. Por el contrario, en el algoritmo PA, ésta se hace teniendo en cuenta que, dado un patrón $x \rightarrow z \leftarrow y$, la instanciación del nodo z debe incrementar el grado de dependencia de x e y. Con respecto al patrón $x \leftarrow z \rightarrow y$, o cualquiera de los otros dos existentes ($x \leftarrow z \leftarrow y$ o $x \rightarrow z \rightarrow y$), pasa lo contrario, es decir, la instanciación de z hace que las otras dos variables sean más independientes. Por tanto, la idea para orientar es comparar el grado de dependencia entre x e y antes de la instanciación de z con el grado de dependencia después de la misma y dirigiendo los arcos hacia z si el primero es mayor que el segundo. El conjunto de pasos que da el algoritmo PA quedan reflejados en el algoritmo 1.3, con la característica principal de que recupera el poliárbol exacto asociado a un modelo de dependencia isomorfo a un poliárbol.

Algoritmo 1.3 Aprendizaje de un poliárbol (Método PA).

```
Generar el árbol generador de costo máximo utilizando el algoritmo de Chow y Liu pero
calculando el peso de la arista como el producto de la dependencia marginal de los dos
nodos que une y la condicionada con respecto a cada una de las otras variables.
para Cada tripleta de nodos tal que x − z − y en el árbol hacer
  si La dependencia marginal de x e y es mayor que la condicionada sobre x entonces
    Orientar el subgrafo x − z − y como x → z ← y
  fin si
fin para
Dirigir las aristas restantes sin introducir ninguna conexión cabeza-cabeza.
```

1.3. Aplicación de las redes bayesianas a la recuperación de información.

El proceso completo de recuperación de información está caracterizado por una propiedad básica, la incertidumbre, ya que cada una de las etapas que lo compone poseen características especiales que permiten calificarlo como un proceso incierto [TC97]:

- La consulta formulada por el usuario no es más que una descripción vaga de la necesidad de información que éste tiene en su mente, ya que debido a ciertos motivos el usuario no puede conseguir una representación fiel de la misma.

- La creación de las representaciones de los documentos y consultas es otro ejemplo, ya que da como resultado una caracterización incompleta, en forma de lista de términos, del contenido de los documentos y de las consultas.
- La etapa de asignación del grado de relevancia de un documento con respecto a una consulta también está influenciada por la incertidumbre, en este caso doble: por un lado la que proviene de los dos anteriores puntos, y por otro, debido a las múltiples representaciones que puede poseer un mismo concepto y a que dichos conceptos no son independientes entre sí.

Basadas, como hemos visto, en métodos probabilísticos, las redes bayesianas han demostrado ser unas herramientas excelentes para manejar la incertidumbre, incluso en el campo de la recuperación de información donde han sido aplicadas de manera exitosa como una extensión del modelo de recuperación probabilístico.

Son ya muchos los trabajos que se han desarrollado en este campo. Vamos a realizar seguidamente una breve revisión de los mismos, clasificándolos según el tipo de aplicación:

- Modelos de recuperación.

  Los pioneros en aplicar las redes bayesianas a la recuperación fueron Turtle y Croft [Tur90, TC91, TC91b, TC92, CT92, Bro94, Gre96, GCT97], desarrollando el modelo denominado *Inference Network Model*, que posteriormente dio lugar al S.R.I. *INQUERY* [CCH92, CCB95]. Este modelo está constituido por dos redes bayesianas: la de documentos y la de consultas. La primera, fija en todo el proceso, está compuesta básicamente por dos tipos de nodos, los cuales representan los documentos y los términos de la colección. Los documentos se conectan con los términos con los cuales han sido indexados, mediante arcos apuntando a los nodos término. La red de consultas representa una necesidad de información, expresada por diferentes consultas que atienden a posibles representaciones distintas, cada una de ellas asociada a los denominados nodos consulta. Ambas redes se unen con arcos desde los nodos que representan términos en los documentos hacia los nodos de términos en las consultas. La inferencia se lleva a cabo instanciando cada documento individualmente. Por tanto, para cada evidencia se calcula la probabilidad de que la consulta sea satisfecha, dado que un documento se ha observado en la colección.

  Un trabajo muy relacionado con este anterior es el de Ghazfan y col. En [GIS96, IGS96] proponen modificaciones a la red de Croft y Turtle para dar "una semántica correcta al proceso de inferencia", para lo cual cambian el sentido de los arcos e instancian sólo el nodo consulta en lugar de cada documento, propagando hacia los nodos documentos y sólo una vez.

  El segundo modelo más importante es del de Ribeiro y Reis [RM96, Rei00, SRCMZ00], los cuales desarrollan el denominado *belief network model*. En él, la red bayesiana posee nodos que representan los términos de la colección; nodos documentos, asociados con los documentos y cuyos padres son los nodos términos relacionados con los términos con los que han sido indexados; y, por último, el nodo consulta, conectado con los nodos términos de la misma manera que los documentos. La inferencia se realiza prácticamente mediante la aplicación del teorema de Bayes. A juicio de sus creadores, su modelo presenta un espacio muestral mucho más claro que la red de inferencia, y tiene un fundamento teórico más sólido.

Estos tres modelos anteriores no aprenden las redes sino que "imponen" las topologías directamente a partir de los términos de los documentos. Por el contrario, Fung y col. en [FCAT90, FF93, FF95] han desarrollado también un modelo basado en redes bayesianas, en el cual la diferencia más destacable con los modelos anteriores es que aprenden las redes que utilizan, estableciendo una base de conocimiento a partir de las relaciones entre los términos de la colección.

Un cuarto modelo es el de Bruza y col. [BI94, BG94b, Ijd94, IBH95], quienes trabajan con *redes de creencia de expresiones índice* (*Index Expression Belief Network*), un conjunto de términos unidos por conectores (and, or, etc.), representando cada una el contenido de un documento en una colección. Estos investigadores construyen una red de creencia con dichas expresiones, incluidas las pertenecientes a la consulta, las cuales serán las evidencias de la red.

- Realimentación de la relevancia y expansión de consultas.

  Los dos modelos principales (Inference Network y Belief Network) han sido adaptados para realizar realimentación de relevancia [HC93, Rei00] como forma para aumentar el rendimiento de la recuperación, al igual que para llevar a cabo sólo expansión [HFC94]. También se han empleado para construir tesauros [JC94]. Por otro lado, Park y col. [PC96] construyen un tesauro basado en una red bayesiana sigmoide, utilizada para poner en práctica un proceso de expansión de consultas. Por último, en [HH98] se usan para modelar relaciones entre las palabras pertenecientes a una consulta.

- Agrupamiento y clasificación de documentos.

  El agrupamiento de documentos y la clasificación son dos de las áreas de la recuperación de información en las que se ha estado trabajando desde los principios de esta disciplina, pero donde no se habían aplicado las técnicas existentes en el ámbito de las redes bayesianas. Han sido Sahami y col. [Sah96, Sah98, SDHH98, SYB98, DPHS98] los que las han puesto en práctica, desarrollando a la vez el S.R.I. *SONIA*, que posee módulos para realizar estas tareas.

- Aplicaciones al hipertexto.

  Además del modelo de Croft y Turtle [CT93], dos son los enfoques que aplican las redes bayesianas al hipertexto. Estos son muy similares y fueron desarrollados por Frisse y Cousins [FC89] y Savoy y Desbois [SD91, Sav93, Sav95]. Ambos utilizan una red con estructura de árbol cuyos nodos son términos en un contexto de hipertexto. En el primer enfoque, el árbol tiene una estructura predefinida, mientras que en el segundo se aprende mediante el algoritmo del "árbol generador maximal". Otra diferencia radica en la construcción de las matrices de probabilidades: Savoy y col. usan una metodología bien fundamentada en el campo de la recuperación de información, mientras que Frisse y col. la estiman sin utilizar métodos numéricos. En ambos casos el árbol se utiliza para recalcular los pesos de los términos de los documentos con la probabilidad a posteriori, obtenidos tras instanciar los nodos de la consulta y posteriormente comenzar la navegación en un sistema hipertexto.

- Otras aplicaciones.

  Para finalizar este repaso, nos queda citar el trabajo de Tzeras y Hartmann [TH93], los cuales presentan un modelo de indexación donde la estructura básica es una red bayesiana, y la aplicación de INQUERY a la búsqueda de documentos en colecciones distribuidas [CLC95].

1.3.1. Descripción de los principales modelos de recuperación basados en redes bayesianas.

En este apartado vamos a describir brevemente los tres principales modelos de recuperación basados en redes bayesianas, estableciendo las características más importantes de los mismos y sus diferencias principales. Nos centraremos en la topología de la red subyacente al modelo, para pasar seguidamente a explicar cómo hacen la estimación de las probabilidades. Finalmente, comentaremos los mecanismos con los cuales llevan a cabo la inferencia.

Comencemos inicialmente comentando la estructura de red del modelo de Croft y Turtle, planteada en [Tur90] en dos partes, como se puede observar en la figura 1.10: la red de documentos y la de consultas. La primera representa la colección documental, se construye sólo una vez y su estructura no cambia en ningún momento. Está compuesta por los siguientes tipos de variables (todos los nodos de la red son variables aleatorias binarias que toman los valores *{verdadero, falso}*), siguiendo la notación original:

- *Nodos documento* (d_i): representan los documentos de la colección y corresponden con el suceso de que un documento específico ha sido observado.

- *Nodos de representación de textos* (t_j): como un mismo documento puede representarse de diferentes maneras en este modelo, estos nodos indican la manera en que se ha llevado a cabo dicha representación, estando asociados al suceso de que una representación ha sido observada.

- *Nodos de representación de conceptos* (r_k): son los términos con los que se ha indexado los documentos, aunque dependen de las representaciones utilizadas para cada documento, indicando que una representación ha sido observada.

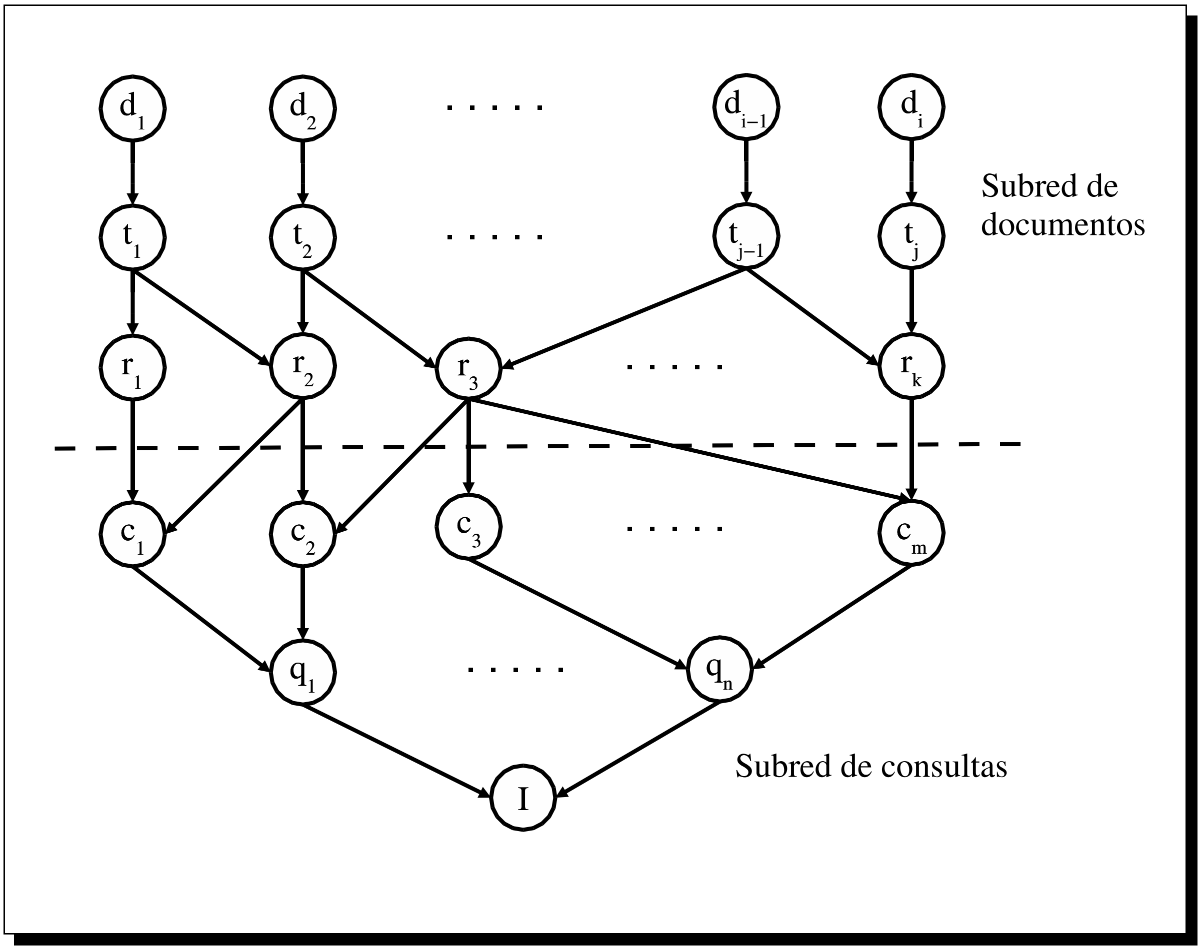


Figura 1.10.: Topología del modelo Inference Network.

De los nodos documentos surgirán arcos hacia los nodos de representación de textos asociados con ellos, y de éstos a los de representación de conceptos (un arco por cada uno de los términos que hayan sido asociados al documento). Con esta orientación, los autores del modelo indican que la observación de un documento es la causa del incremento de creencia en las variables asociadas con los términos que lo indexan, es decir, los nodos de representación de conceptos.

Por otro lado, la red de consultas contiene los tres tipos de nodos que a continuación se indican:

- Un único nodo *necesidad de información* (I), que representa la necesidad de información interna del usuario, estando asociado al evento: se ha encontrado una necesidad de información.
- Nodos de representación de consulta (q), cuya existencia se basa en el hecho de que una necesidad de información puede expresarse mediante varias consultas a la vez. Cada una

de estas consultas quedará plasmada en la red de consultas por un nodo de este tipo. Por ejemplo, una consulta formulada a base de citar los términos que la componen (estilo vectorial) se puede combinar con otra formulada en términos booleanos.

- Nodos conceptos de la consulta (c_m), que representan los términos con los que se ha indexado la consulta.

En este caso, al nodo necesidad de información llegará un arco de cada uno de los nodos de representación de consultas existentes, y a éstos los procedentes de conceptos contenidos en la consulta.

Ambas redes se unirán por medio de los nodos de representación de conceptos y los de concepto de la consultas: los segundos tendrán como padres a los primeros (un nodo concepto de consulta tiene como padre a uno o varios nodos de representación de conceptos, según el número de representaciones que se hayan hecho para un mismo documento).

Esta sería la estructura más genérica, aunque para situaciones en donde solo exista una única forma de representación, habría una versión más simplificada en donde la red quedará compuesta por un nodo necesidad de información, los nodos de representación de consultas, los nodos de representación de conceptos y los nodos documento, como se puede ver en la figura 1.11.

Croft y Turtle justifican la orientación de los arcos en sentido descendente por razones relacionadas puramente con la causalidad: de las causas a los efectos. Así, mantienen que la observación de un documento aumenta nuestra creencia en una representación del texto, que a su vez incrementa la creencia en el conjunto de nodos de representación de conceptos, éstos en los nodos de conceptos de la consulta, que fortalece la crencia de los nodos de representación de la consulta, y finalmente en el nodo de la necesidad de información.

Por el contrario, Ghazfan y col. [GIS96, IGS96] no piensan así y, manteniendo los tipos de nodos de la versión simplificada de Croft y Turtle (excepto los nodos de consulta que desaparecen), cambian la orientación de los arcos de la red, siendo la raíz en su modelo el nodo de la necesidad de información y los nodos hojas los documentos. Lo justifican argumentando que la propagación de probabilidades toma una semántica correcta con esa direccionalidad de los arcos. En la figura 1.12, y siguiendo su misma notación, podemos ver los nodos etiquetados como A, B, C y D que representan términos, Q representa la consulta y 1 y 2 son dos documentos.

Por último, Ribeiro y Reis, en su modelo *belief network model* [RM96, Rei00, SRCMZ00], mantienen tres tipos de variables binarias. En la figura 1.13 se aprecia la composición de la red bayesiana: nodos que representan los documentos (D_j), los términos (K_i) y la consulta (Q). Los primeros tienen como padres a los nodos término que están asociados con los términos por los que han sido indexados. Análogamente, el nodo consulta, tiene como padres los nodos términos referentes a los términos existentes en la consulta. Así, los nodos término son las raíces y los nodos documento y consulta son las hojas del grafo. Básicamente toman esta opción porque tratan conceptualmente a las consultas y a los documentos como objetos del mismo tipo.

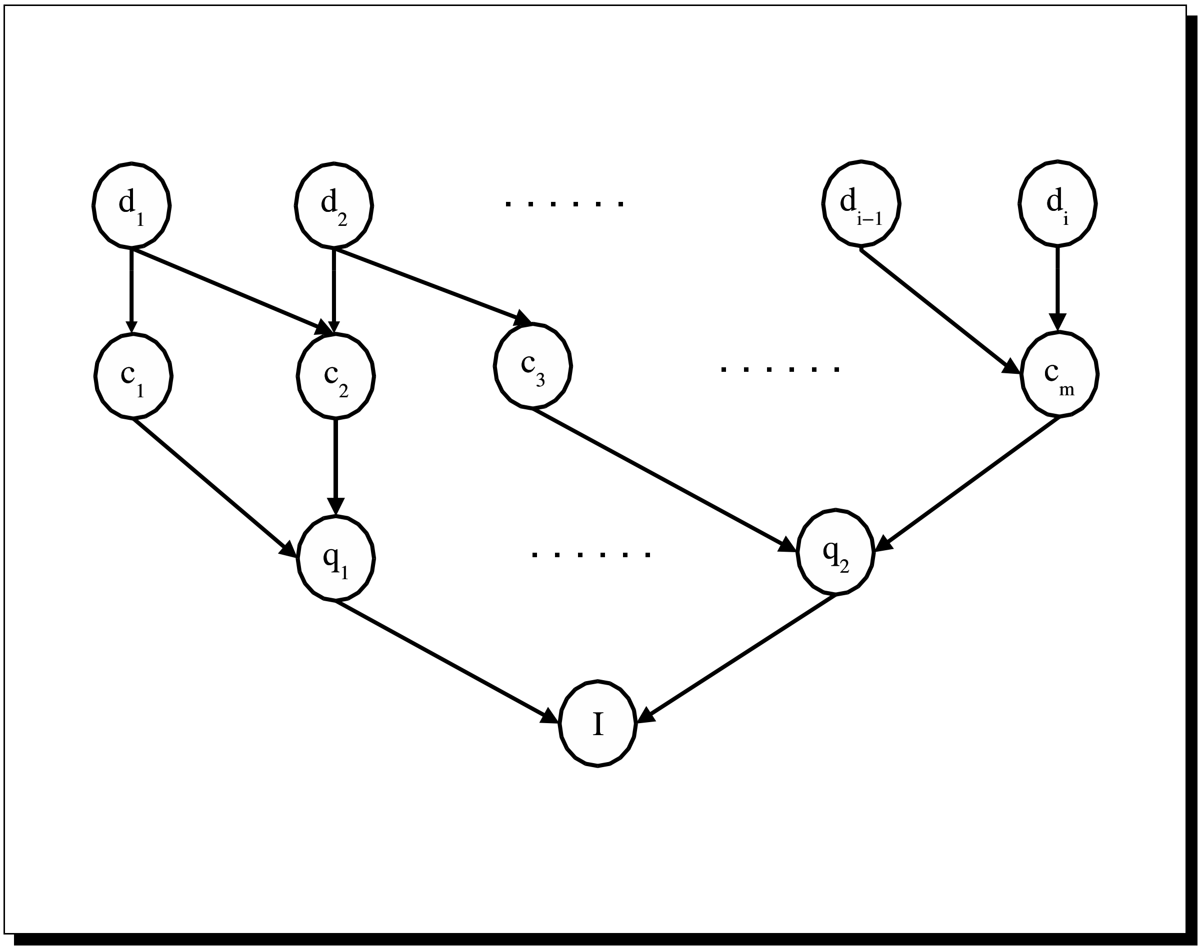


Figura 1.11.: Topología del modelo Inference Network reducido.

En este modelo, sus autores pretenden establecer un marco general en donde se simulen, mediante una estructura de red bayesiana, los modelos clásicos más el modelo *Inference Network*.

En cuanto a la información cuantitativa que albergan cada uno de los tres tipos de redes, el *Inference Network*, en los nodos representación de conceptos se almacenan las siguientes probabilidades:

$$p(r_i = verdad \mid d_j = verdadero) = 0{,}5 + (0{,}5 \cdot ntf_{ij} \cdot nidf_i)$$
$$p(r_i = verdad \mid \text{ todos los padres a falso}) = 0$$

donde ntf_{ij} y $nidf_i$ son las correspondientes normalizaciones, según su propia notación, del tf y del idf respectivamente. El hecho de que no contemplen todas las combinaciones posibles de valores que pueden tomar los padres de un nodo es debido a la forma en que realizan la propagación y que más tarde describiremos.

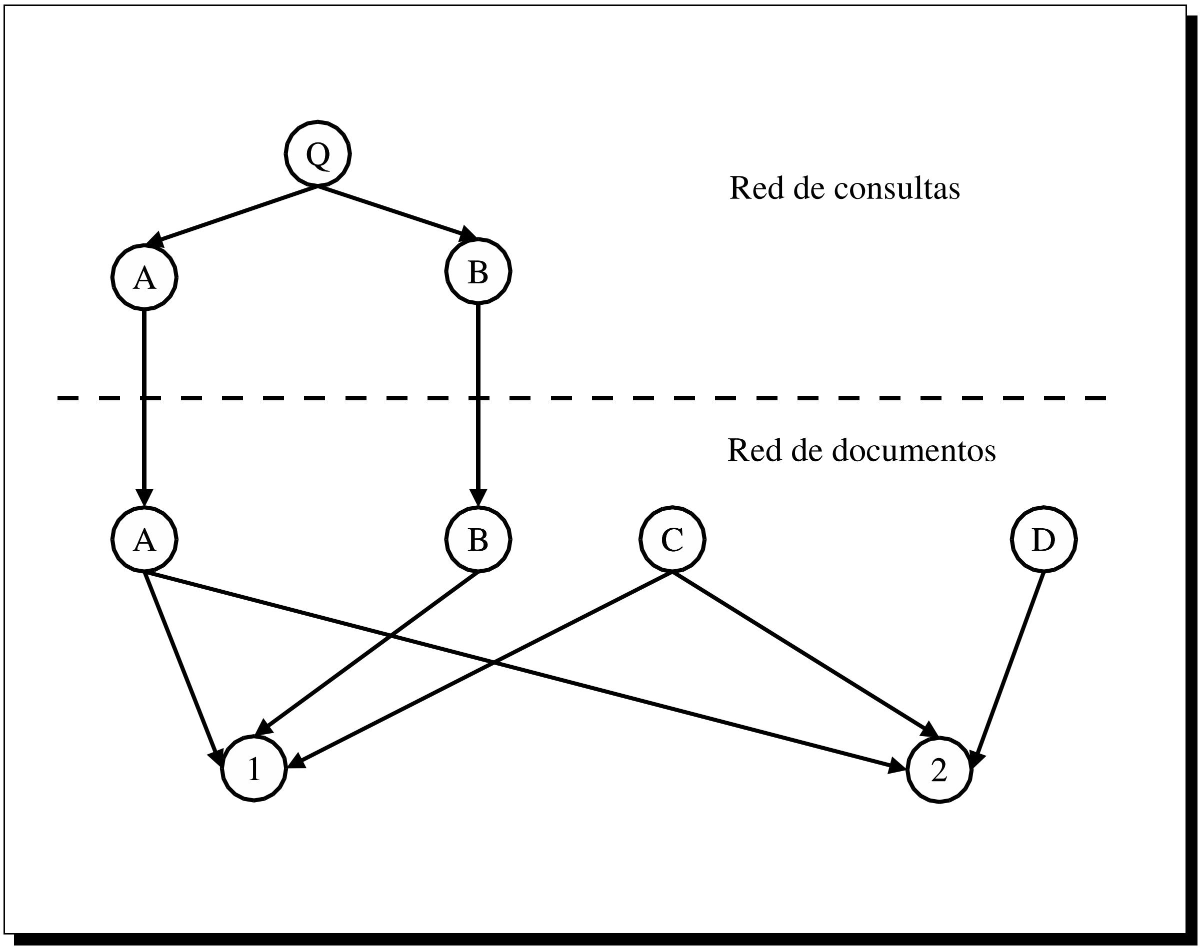


Figura 1.12.: Topología del modelo de Ghazfan y col..

Por otro lado, con respecto al nodo consulta Q, dependiendo del tipo de consulta, se hace una estimación diferente. Si ésta es booleana, se generan aplicando las reglas con las que trabajan los operadores booleanos; en caso de que sea probabilística, se obtiene cada elemento de la matriz como una suma de probabilidades a priori de cada término padre y ponderadas por su peso correspondiente, normalizada toda ella por la suma total de pesos.

En el modelo *Belief network*, no se establecen asignaciones explícitas para cada uno de los tipos de nodos que figuran en su red, sino que, dependiendo del modelo que deseen simular, establecen una forma de cálculo diferente cuyo objetivo es conseguir la misma función de ordenación que el modelo simulado.

En cuanto al modelo de Ghazfan y col., no podemos exponer cómo hacen las estimaciones, pues no lo explicitan en los trabajos de que disponemos.

Seguidamente vamos a centrarnos en cómo llevan a cabo el proceso de inferencia cada uno de los tres modelos anteriores. Comenzando por el modelo Inference Network, el uso que se hace de la red pasa por determinar el grado con el que el apoyo evidencial de un documento

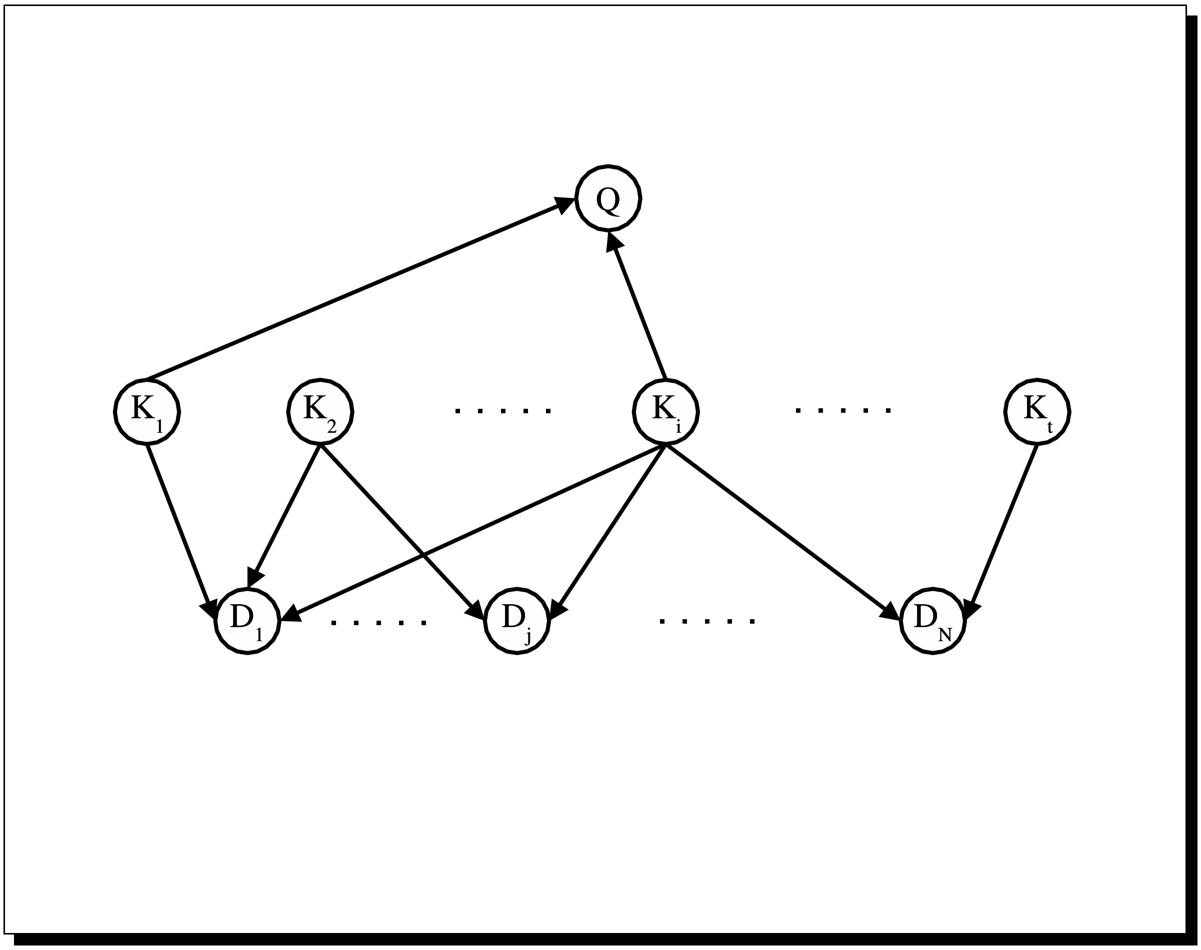


Figura 1.13.: Topología del modelo Belief network.

influye sobre la necesidad de información. Esta medida será la utilizada finalmente para establecer la ordenación de documentos. Este grado se consigue mediante la estimación, una vez instanciado a verdadero, y de manera sucesiva, cada documento de la colección (dejando el resto a falso), de la probabilidad de que la consulta sea relevante dado el documento instanciado. Estos cálculos no se llevan a cabo mediante la ejecución de un método de propagación, sino, debido a la topología de la red, simplemente mediante la evaluación de una función que calcula la probabilidad requerida.

En cuanto al modelo de Ribeiro y col., la inferencia se hace instanciando la consulta y propagando a los documentos, obteniendo la probabilidad de que cada documento sea relevante dado que la consulta también lo es, justo al contrario que el modelo anterior. De esta manera, sólo se tiene que poner en práctica la inferencia una única vez, en oposición al modelo *Inference Network*, que debería hacer tantas propagaciones como documentos (instancias) hubiera. Al igual que en el modelo anterior, no ejecutan ningún algoritmo de propagación, sino que evalúan una función que devuelve la probabilidad condicional buscada para cada documento.

Por último, el modelo de Ghazfan y col., a pesar de no tener como estructura subyacente de la red un árbol, utilizan el método de propagación exacta de Pearl [Pea88], pero modificándolo con objeto de que los mensajes que fluyan a través de la red no estén circulando indefinidamente en los ciclos existentes. Comentar, por último, que ellos también instancian la consulta y propagan hacia los documentos

2. Creación de un tesauro basado en una red bayesiana para expansión de consultas.

Los S.R.I. pueden mejorar su efectividad, es decir, incrementar el número de documentos relevantes recuperados, utilizando un proceso de expansión de consultas, el cual añade automáticamente nuevos términos a la consulta original formulada por un usuario. En este capítulo desarrollamos un método para expandir consultas basado en redes bayesianas. Mediante un algoritmo de aprendizaje, y partiendo de un fichero invertido de una colección, construimos una red bayesiana que representa algunas de las relaciones existentes entre los términos que aparecen en la colección de documentos. Esta red se usa como un tesauro (específico para esa colección) para seleccionar los términos del mismo que más relación tengan con los términos que figuran en la consulta. De esta forma, la consulta original se enriquece con nuevos términos que ayudan a recuperar documentos relevantes, que solamente con la consulta original no podrían obtenerse. Finalmente, presentamos los resultados que obtenemos utilizando este subsistema de expansión como ayuda al S.R.I. *SMART* en tres colecciones de prueba.

2.1. Introducción a los tesauros y a la expansión de consultas.

Podríamos definir de forma genérica el concepto de *tesauro* como un conjunto de palabras o frases pertenecientes a un dominio específico, tal que cada una de ellas contiene a su vez un conjunto de otras palabras o frases con las que está relacionado [BR99]: sinónimos, antónimos, términos más generales y específicos, y otros con los que están estrechamente relacionadas [Kor97], pero que no se pueden enmarcar en estos tipos de relaciones anteriores.

Los tesauros tienen varias finalidades, pero las dos principales son las siguientes:

- Ayudar al indexador a realizar su labor, seleccionando los términos del tesauro que mejor representen el contenido del documento.

- Servir como herramienta para que el usuario pueda diseñar consultas más cercanas a su necesidad de información y establecer una jerarquía con la que poder hacer más general o específica una consulta atendiendo a los términos que contiene.

Así, por ejemplo, si el S.R.I. no recupera suficientes documentos relevantes, el tesauro puede ser de utilidad para expandir la consulta y, por tanto, añadir términos con los que conseguir más documentos interesantes. Por el contrario, si éste devuelve un material muy numeroso, esta herramienta sugerirá un vocabulario más específico que producirá una focalización de la consulta [Sri92].

Según Jing y Croft [JC94], tres son las grandes áreas que centran el trabajo con respecto a los tesauros:

- La construcción del tesauro.
- El acceso al tesauro.
- La evaluación de la calidad del tesauro.

En cuanto a la primera tarea, los tesauros pueden ser construidos manual o automáticamente [Sri92]. Los primeros se basan en el trabajo del diseñador, el cual debe recopilar todo el material susceptible de ser introducido en el tesauro y organizarlo, identificando las relaciones entre ellos. Los automáticos delegan en un programa la tarea de la construcción de los tesauros. Existen numerosos enfoques que tratan este problema (Schütze y Pedersen hacen una revisión de las principales líneas en la construcción automática de tesauros en [SP97]), de entre los cuales nos vamos a centrar en los que se basan en técnicas estadísticas para establecer las principales relaciones entre los términos, ya que el método que presentamos en este capítulo se clasifica en este grupo.

Esta clase de métodos de construcción de tesauros se basa en la *hipótesis de asociación*, la cual fue establecida por van Rijsbergen [Rij79] y dice:

> *Si un término es bueno discriminando documentos relevantes de los que no lo son, entonces es probable que cualquier otro término relacionado estrechamente con él sea también bueno para esta tarea.*

Así, la forma en que se pone en práctica la hipótesis anterior es: partiendo de todos los términos con los que se creará el tesauro, se calcula alguna medida de similitud entre ellos, y posteriormente y teniendo ésta en cuenta, se organiza el vocabulario en una estructura jerárquica. Por ejemplo, Salton y McGill en [SM83], parten del fichero invertido de una colección y calculan una medida de similitud entre cada par de términos, basada en la suma del producto de los pesos de los términos en los documentos en que coinciden. Seguidamente aplican el método conocido como *single-link* [Rij79], para crear una jerarquía de clases de términos. En esta línea, cabe destacar también el trabajo llevado a cabo por van Rijsbergen en la generación de un árbol de dependencia [Rij79] utilizado con el modelo probabilístico, el cual será estudiado con más detalle posteriormente.

La construcción del tesauro se puede realizar con todos los términos de la colección, en cuyo caso diremos que se ha empleado una *estrategia global*. Alternativamente, tendremos una *estrategia local*, cuando se genera con los términos de los documentos que se juzgan como relevantes por parte del usuario [BR99], utilizándose en este caso, para expandir consultas, como veremos a continuación.

Claramente, los tesauros generados con estos métodos son totalmente dependientes de la colección con la que se esté trabajando y, por tanto, sólo pueden ser usados en los contextos para los que han sido elaborados. Además, estos enfoques son capaces de determinar que dos términos están relacionados entre sí, pero no establecen el tipo de relación real que hay entre ellos, con los problemas semánticos que pueden ocasionarse por esta ambigüedad en las relaciones. Por último, también se les critica por el hecho de no utilizar ningún tipo de limitación relacionada con la distancia, medida en número de palabras, que separa la ocurrencia de dos términos en el texto, ya que tienen en cuenta el texto completo y eso puede dar lugar a que las relaciones entre términos no sean totalmente reales [JC94]. Obviamente, los nuevos métodos de construcción de tesauros van solucionando estos problemas.

En cuanto al uso que se hace de los tesauros, es decir, para qué se utiliza la información que contienen, el principal es aquél que tiene por objeto mejorar la calidad de la consulta. De entre esos métodos, destacamos el de la *expansión de la consulta*, el cual intenta obtener más documentos relevantes mediante la inclusión en la consulta original de nuevos términos. Se pretende que estos nuevos términos, relacionados de alguna forma con los originales, consigan recuperar aquellos documentos relevantes que no pueden ser conseguidos sólo con los términos de la consulta original. Han y col. hacen en [HFC94] una clasificación en seis grupos de las diferentes técnicas de expansión de consulta, de entre las cuales nos quedaremos con aquella basada en el empleo de tesauros construidos estadísticamente. Este tipo de expansión tiene su justificación, al igual que la propia construcción, en la hipótesis de asociación, ya comentada anteriormente.

La selección de términos, normalmente, se hace tomando, para cada término de la consulta, aquellos otros términos del tesauro que estén más relacionados con él [SM83]. Esta forma de selección "local" tiene como principal inconveniente que pierde la información suministrada por el sentido global de la consulta [JC94].

Otro asunto importante es el número de términos con los que se expande la consulta. Buckley y col. aseguran en [BSAS94] que la efectividad de la expansión aumenta linealmente con el logaritmo del número de términos añadidos, hasta un punto en que comienza a disminuir. Por otro lado, Harman, en [Har92, Har92b], defiende el hecho empírico de expandir con un número de 20 términos aproximadamente (valor totalmente dependiente de la colección usada).

Por último, comentar que la expansión de consultas se puede poner en práctica de dos formas completamente diferentes: por un lado, aplicada a la consulta original antes de que se haya recuperado con ella, como complemento a la misma; y por otro, como uno de los componentes de la técnica de realimentación de relevancia, es decir, una vez que se ha recuperado con la consulta original, para recuperar más documentos relevantes distintos a los ya obtenidos.

2.2. Enfoques basados en redes bayesianas para la construcción y uso de tesauros.

Una vez realizada la introducción de la sección anterior a los conceptos básicos de construcción de tesauros y expansión de consultas, vamos a presentar las aplicaciones de las redes bayesianas en este campo de la recuperación de información. En esta sección vamos a revisarlas brevemente, y en la siguiente estableceremos las diferencias fundamentales con nuestro enfoque, una vez que haya sido expuesto.

Comencemos por el trabajo de van Rijsbergen y col. [Rij77, Rij79, HR78, RHP81] que, aunque no utilizan como tal una red bayesiana, construyen un *árbol generador de peso máximo*[1] *(A.G.P.M.)*, también conocido como *Maximum Spanning Tree (M.S.T.)* [Chr75], para realizar expansión de consultas en el modelo probabilístico. Inicialmente, este modelo se formuló suponiendo que los términos de una colección son independientes entre sí. Esta suposición es muy estricta y poco real, con lo que en este modelo se buscó una forma de determinar las principales dependencias existentes entre los términos [Rij77, HR78, Rij79] que aparecían en los conjuntos de documentos relevantes y no relevantes (estrategia local): la construcción de un árbol generador de peso máximo para cada conjunto, basada en la *entropía cruzada de Kullback - Leibler* (expresión (2.2)) y en el método de Chow y Liu [CL68]. La distribución de probabilidad conjunta obtenida a partir de cada árbol servirá posteriormente para evaluar la función de ponderación para cada documento.

Basado también en el modelo probabilístico, en [RHP81] se presenta un método fundado igualmente en el árbol generador de peso máximo, aunque en este caso construido con diferentes medidas de asociación, para llevar a cabo expansión de consultas en este esquema de recuperación. A diferencia de la aproximación anterior, sólo se construye un único árbol a partir de la información de coocurrencias de términos en toda la colección; es decir, se aplica un enfoque global. Dada una consulta, ésta se expande mediante la inclusión en ella de los términos del árbol que están conectados con cada uno de los términos de la consulta. En ambos casos la selección de términos se hace localmente a cada término de la consulta original.

El primer modelo que plantea el uso de una red bayesiana para llevar a cabo la expansión, entre otras herramientas básicas desarrolladas para facilitar la navegación en los entornos hipertexto donde se enmarca este trabajo, es el presentado por Frisse y Cousins en [FC89]. En él, la estructura de la red que almacena términos médicos es también un árbol, y viene impuesta manualmente [Fri88], y las matrices de probabilidad se estiman mediante métodos empíricos, funciones heurísticas y observaciones de los intereses de los lectores.

Un enfoque muy parecido es el introducido por Savoy y col. en [SD91, Sav92], también encuadrado en sistemas hipertexto, aunque con dos diferencias principales que lo hacen mucho más versátil con respecto al anterior:

- La red bayesiana compuesta por los términos pertenecientes a los documentos de hiper-

[1] Un árbol generador de peso máximo es un árbol obtenido a partir de un grafo cualquiera, en el que la suma de pesos de sus aristas es máxima.

texto es aprendida mediante el método del árbol generador de peso máximo, a partir del glosario completo de la colección.

- La estimación de las matrices de probabilidad almacenadas en el árbol se basa en la fórmula de Jaccard [Rij79].

Ambas aplicaciones usan el algoritmo de propagación en árboles de Pearl [Pea88] para, una vez instanciados los términos de la consulta, ofrecerle al usuario un conjunto de nodos del grafo de documentos a visitar dada una posición inicial en él. Además, realizan una selección global a partir de la consulta completa.

El modelo de Croft y Turtle también se ha utilizado como tesauro para posteriormente expandir consultas. Ya en su tesis doctoral, Turtle hace referencia a la extensión de la red de documentos con relaciones de sinonimia entre los términos de la red de documentos y generalidad o especificidad [Tur90]. Otra aplicación de este modelo fue el trabajo de Jing y Croft [JC94], los cuales construyen un tesauro basado en las coocurrencias de frases y términos, incorporándolo posteriormente a la red de inferencia del S.R.I. *INQUERY* [CCH92], como si fueran documentos (cada entrada del tesauro se representa como una lista de términos con los que está relacionado). Así, una vez que se ha efectuado una consulta, INQUERY devuelve el conjunto de términos más relevantes con respecto a ella, siendo utilizados posteriormente para expandir la consulta original. Esta misma idea se explota en [HFC94], para realizar expansión de consultas en textos en japonés, y en otros trabajos como [XC96, XC00].

Aunque en otro contexto, Sahami [Sah98] también aprende un árbol generador de peso máximo. En este caso, lo ha utilizado como la base para llevar a cabo un proceso de selección de características y así reducir las dimensiones del problema del agrupamiento de documentos.

Por último, comentar el trabajo de los autores Park y Choi [PC96], los cuales construyen una variedad de red bayesiana, denominada *sigmoide*, a partir de las similitudes entre los términos, empleándola también posteriormente para realizar la expansión de consultas.

2.3. Construcción del tesauro.

Como ya hemos comentado anteriormente, un tesauro contiene un conjunto de palabras clave y las relaciones existentes entre ellas. Una red bayesiana, a su vez, es capaz de representar las relaciones establecidas entre el conjunto de variables que almacena. Por tanto, podemos utilizar un grafo de este tipo para representar las relaciones de independencia o dependencia existentes entre los términos de una colección documental.

Dada una colección de documentos, la idea que vamos a desarrollar en esta sección es plantear cómo se realizará la construcción de un tesauro a partir de los términos contenidos en dicha colección. Este tesauro será una de las piezas fundamentales para el desarrollo del modelo de recuperación de información basado en redes bayesianas que presentaremos en el siguiente capítulo.

2.3.1. El algoritmo de aprendizaje de la red bayesiana.

Un primer problema que se plantea es elegir la topología subyacente de la red bayesiana que se aprenderá. Para realizar una elección correcta hay que tener en cuenta que en el campo de la R.I., las bases de datos documentales que se manejan pueden llegar a tener decenas de miles de documentos, lo que implica que el número de términos implicados en la indexación puede ser astronómico. Está claro que cuanto más compleja sea esta topología más fielmente reproducirá las relaciones reales de independencia y dependencia existentes entre los términos de la colección, pero a su vez, y teniendo en cuenta el número de términos manejados, el proceso de aprendizaje y, posteriormente, su uso (la propagación de probabilidades) serán tareas muy costosas. Por otro lado, con grafos más simples, es evidente que se pierde en precisión pero, en contraposición, para ellos existe un conjunto de algoritmos de aprendizaje y propagación muy eficientes [Pea88] que hacen que sea totalmente factible el uso de estas estructuras dentro del contexto donde nos movemos.

De entre estas topologías menos complejas, destacamos los grafos conocidos como *grafos simplemente conectados* [Chr75] en los cuales entre dos vértices cualesquiera existe a lo sumo un único camino que los une. La elección de este tipo de grafos como base para sustentar el tesauro se ha llevado a cabo por dos razones, ya esbozadas anteriormente:

- Como cada término tiene asociado en la red un nodo, como ahora explicaremos, nuestro grafo sería muy grande. Por tanto, la estimación de la red a partir de datos empíricos, es decir, el aprendizaje, sería un proceso que consumiría mucho tiempo. Esta tarea puede ser aliviada aplicando algoritmos de aprendizaje eficientes [Cam98, RP89] que construyen grafos más simples, como son las redes simplemente conectadas, consideradas como aproximaciones de modelos más complejos debido a la pérdida de expresividad (ya que las relaciones de dependencia e independencia que las redes simplemente conectadas pueden representar son más restrictivas que para grafos generales).
- La segunda razón está basada en los métodos de inferencia disponibles para redes bayesianas, es decir, los métodos de propagación. De nuevo tratamos con procesos que consumen mucho tiempo, aunque para redes simplemente conectadas existen métodos exactos y eficientes que se ejecutan en un tiempo proporcional al número de nodos existentes [Pea88].

Ejemplos de estos grafos simplemente conectados son los árboles y poliárboles. De entre estos dos tipos de grafos simples, hemos elegido los *poliárboles* (grafos en los que no existe más de un camino dirigido conectando cada par de nodos), por ser los más completos de este grupo.

Una vez determinada la topología de la red bayesiana subyacente del tesauro, vamos a pasar a describir cómo se construye la red a partir de los datos empíricos, proceso que se conoce como *aprendizaje*.

Este proceso usa como fichero de aprendizaje un fichero invertido. La estructura de este fichero está compuesta por tantas líneas como términos existan en la colección, conteniendo cada una de ellas los documentos en los que aparece dicho término. Así, cada término T_i del glosario de la colección, $\mathcal{T}$, tiene asociado en la red bayesiana una variable aleatoria binaria [2], que notaremos igual que el término, T_i, la cual podrá tomar uno de los siguientes valores: $T_i \in \{t_i, \bar{t}_i\}$, donde t_i significa "*el término T_i es relevante*", y $\bar{t}_i$, "*el término T_i no es relevante*". Una vez especificado el contenido de la red, pasamos a describir el algoritmo que se ha diseñado para aprender el poliárbol de términos, que se puede ver en el Algoritmo 2.1.

Algoritmo 2.1 Aprendizaje del poliárbol de términos.

1: $G \leftarrow \emptyset$
2: **para** cada par de nodos $T_i, T_j \in \mathcal{T}$ **hacer**
3:     Calcular $Dep(T_i, T_j \mid \emptyset)$.
4: **fin para**
5: Construir el árbol generador maximal, G, en el que el peso de cada arista T_i–T_j es:

$$Dep(T_i, T_j) = \begin{cases} Dep(T_i, T_j \mid \emptyset) & \text{si } \neg I(T_i, T_j \mid \emptyset) \\ 0 & \text{si } I(T_i, T_j \mid \emptyset) \end{cases} \tag{2.1}$$

6: **para** cada tripleta de nodos $T_i, T_j, T_k \in \mathcal{T}$, tales que T_i–T_k, T_k–$T_j \in G$ **hacer**
7:     **si** $Dep(T_i, T_j \mid \emptyset) < Dep(T_i, T_j \mid T_k)$ y $\neg I(T_i, T_j | T_k)$ **entonces**
8:         Dirigir el subgrafo T_i–T_k–T_j como $T_i \rightarrow T_k \leftarrow T_j$.
9:     **fin si**
10: **fin para**
11: Dirigir los vértices restantes sin introducir ninguna conexión cabeza–cabeza.
12: Devolver G.

Este algoritmo es muy parecido a dos algoritmos de aprendizaje de poliárboles: *el algoritmo PA* [Cam98] y el de Rebane y Pearl [RP89], que notaremos como algoritmo *RP* (ambos algoritmos utilizan el método de Chow y Liu para construir árboles de dependencias [CL68]). Realmente podríamos decir que es una combinación de ambos algoritmos con algunas características adicionales.

Como puede verse, el algoritmo se compone de tres partes principales:

- En el paso segundo, tras iniciar el poliárbol al vacío, calculamos los grados de dependencia entre todos los pares de nodos.
- La segunda parte, correspondiente al paso quinto, representa la construcción del esqueleto (árbol generador de peso máximo) y
- la última parte, desde el paso sexto hasta el final, lleva a cabo la orientación de las aristas del árbol, creando finalmente un poliárbol.

[2]Hablaremos indistintamente de variable aleatoria binaria o nodo cuando nos estemos refiriendo a los vértices de un grafo. Incluso, por simplicidad, haremos referencia a *términos del poliárbol* en lugar de a *los nodos del poliárbol relacionados con los términos del glosario*.

Se pueden hacer varios comentarios sobre cada una de estas tres partes. Primeramente, la medida utilizada para establecer la dependencia entre dos nodos (la cual es, en cierto sentido, análoga a las funciones empleadas por los modelos de recuperación para medir la similitud entre términos de una colección) es la siguiente:

$$Dep(T_i, T_j \mid \emptyset) = \sum_{\mathbf{T_i}, \mathbf{T_j}} p(\mathbf{T_i}, \mathbf{T_j}) \ln\left(\frac{p(\mathbf{T_i}, \mathbf{T_j})}{p(\mathbf{T_i})p(\mathbf{T_j})}\right) \tag{2.2}$$

Es decir, la *entropía cruzada de Kullback-Leibler* (también conocida como *expected mutual information measure, E.M.I.M.*, es decir, la ecuación (1.20) del capítulo primero) que mide el grado de dependencia entre dos variables T_i y T_j, siendo 0 si son marginalmente independientes, y cuanto más dependientes sean T_i y T_j, mayor será $Dep(T_i, T_j \mid \emptyset)$.

Todas las probabilidades $p(\mathbf{T_i}, \mathbf{T_j})$, que aparecen en la fórmula (2.2) se estiman a partir de un fichero invertido, contando las frecuencias de aparición de los términos. Por ejemplo, $p(t_i, \bar{t}_j)$ es la probabilidad de que el término T_i aparezca en un documento y no lo haga T_j, y $p(t_i)$ es la probabilidad de ocurrencia del término T_i (asumimos que un término T_i es relevante para un documento si aparece en él).

Nosotros utilizamos, como hace el algoritmo RP, la entropía marginal cruzada como medida de dependencia. El algoritmo PA usa una combinación de $Dep(T_i, T_j \mid \emptyset)$ y los grados de dependencia condicionales (medidas de información mutua condicionada) $Dep(T_i, T_j \mid T_k)$, para cualquier otro nodo T_k:

$$Dep(T_i, T_j \mid T_k) = \sum_{\mathbf{T_i}, \mathbf{T_j}, \mathbf{T_k}} p(\mathbf{T_i}, \mathbf{T_j}, \mathbf{T_k}) \ln\left(\frac{p(\mathbf{T_i}, \mathbf{T_j}, \mathbf{T_k})p(\mathbf{T_k})}{p(\mathbf{T_i}, \mathbf{T_k})p(\mathbf{T_j}, \mathbf{T_k})}\right) \tag{2.3}$$

La razón por la cual no hemos combinado la dependencia marginal de dos términos con las dependencias condicionales de dichos términos con el resto de ellos ha sido que, debido a la gran cantidad de palabras clave en una colección, los cálculos de estas dependencias condicionales, aunque sólo se obtengan una única vez, llevan mucho tiempo y además necesitan un espacio en disco bastante amplio para almacenarlas. Estos dos obstáculos anteriores nos han llevado a utilizar sólo la dependencia marginal como medida de dependencia. Están por estudiar métodos que incluyan las dependencias condicionales en ciertos casos.

El siguiente paso del algoritmo es la creación del esqueleto de la red. Si asumimos que los valores de dependencia son aristas ponderadas en un grafo, este algoritmo obtiene un *árbol generador de peso máximo*. Consideramos dos métodos diferentes para obtener el árbol generador, los algoritmos de Kruskal y de Prim [Chr75], aunque finalmente optamos por el segundo. La razón fundamental para esta elección fue que el algoritmo de Kruskal se aconseja para grafos poco densos y no para completos, como es nuestro caso, ya que requiere ordenar todas las dependencias calculadas. El método de Prim comienza con un árbol compuesto por un único nodo, y en cada paso añade al árbol la arista entre un nodo que pertenezca al árbol y otro fuera de él, tal que posean el grado de dependencia máximo, hasta que se añaden $M-1$ aristas, siendo M el número de nodos en el grafo.

Debido al gran número de términos que componen habitualmente una colección, los valores de las dependencias son en general muy bajos, por lo que el algoritmo de aprendizaje en ocasiones no tiene otra opción y elige el valor más alto de todas las dependencias, a pesar de ser un valor muy bajo, añadiendo la correspondiente arista al árbol. El problema yace en el hecho de que, en este caso, los dos nodos conectados son más independientes que dependientes y, por tanto, el modelo que estamos construyendo pierde precisión con respecto al original. Para solucionar este problema, el algoritmo, una vez que ha seleccionado una nueva arista T_i–T_j para añadir al árbol, efectúa un test de independencia entre T_i y T_j (un *test Chi Cuadrado* con un grado de libertad basado en el propio valor de $Dep(T_i,T_j \mid \emptyset)$ [3] [Kul68]). Sólo cuando el test de independencia falla, el algoritmo añade la arista al árbol (en el algoritmo 2.1, se nota como $\neg I(T_i,T_j \mid \emptyset)$). Esto es equivalente a redefinir los pesos de las aristas utilizando la expresión (2.1) del algoritmo en lugar de la ecuación (2.2). De esta forma, y como resultado de este paso, podemos obtener un árbol no conectado, es decir, un bosque.

En la figura 2.1 se representa un posible árbol generador de peso máximo construido tras aplicar el paso quinto del algoritmo 2.1 sobre una colección con catorce términos.

Una vez que el esqueleto se ha construido, la última parte del algoritmo trata de la orientación del árbol, obteniendo como resultado un poliárbol. Este proceso está basado en el algoritmo PA: en un patrón cabeza–cabeza $T_i \rightarrow T_k \leftarrow T_j$, la instanciación del nodo cabeza–cabeza debería incrementar el grado de dependencia entre T_i y T_j, mientras que para un patrón que no sea de este tipo anterior, como es el caso de $T_i \leftarrow T_k \rightarrow T_j$, la instanciación del nodo intermedio debería producir el efecto contrario, es decir, la disminución del grado de dependencia entre T_i y T_j. Así, comparamos el grado de dependencia entre T_i y T_j después de instanciar, $Dep(T_i,T_j \mid T_k)$, con el grado de dependencia antes de dicha instanciación, $Dep(T_i,T_j \mid \emptyset)$, se orientarán las aristas hacia T_k si el primero es mayor que el segundo. Finalmente, el algoritmo orienta las aristas restantes sin introducir nuevas conexiones cabeza–cabeza.

Esta estructura produce, en nuestros experimentos preliminares, configuraciones donde varios nodos tienen un gran número de padres, lo cual origina el tener un gran número de distribuciones de probabilidad condicionada, y como consecuencia, se producen problemas de almacenamiento y fiabilidad en la estimación de las mismas. Por esta razón, hemos restringido la regla que produce conexiones cabeza–cabeza, introduciendo otra condición en el antecedente: queremos estar seguros de que si decidimos incluir una conexión cabeza–cabeza, $T_i \rightarrow T_k \leftarrow T_j$, entonces los nodos T_i y T_j no son condicionalmente independientes dado T_k . Así, comprobamos también esta condición otra vez utilizando un test de independencia Chi Cuadrado basado en el valor $Dep(T_i,T_j|T_k)$ (en este caso con dos grados de libertad).

En la figura 2.2 se puede observar el poliárbol final tras concluir la fase de orientación de enlaces a partir del árbol generador de peso máximo de la figura 2.1.

Este algoritmo que se acaba de presentar para aprender la red bayesiana que soportará el tesauro, presenta las siguientes diferencias con respecto a los enfoques comentados en la sección 2.2:

[3] $2 \cdot N \cdot Dep(X,Y \mid Z)$ se aproxima según una distribución Chi Cuadrado con $\mid Z \mid \cdot (\mid X \mid -1)(\mid Y \mid -1)$ grados de libertad, siendo N el número de datos disponibles.

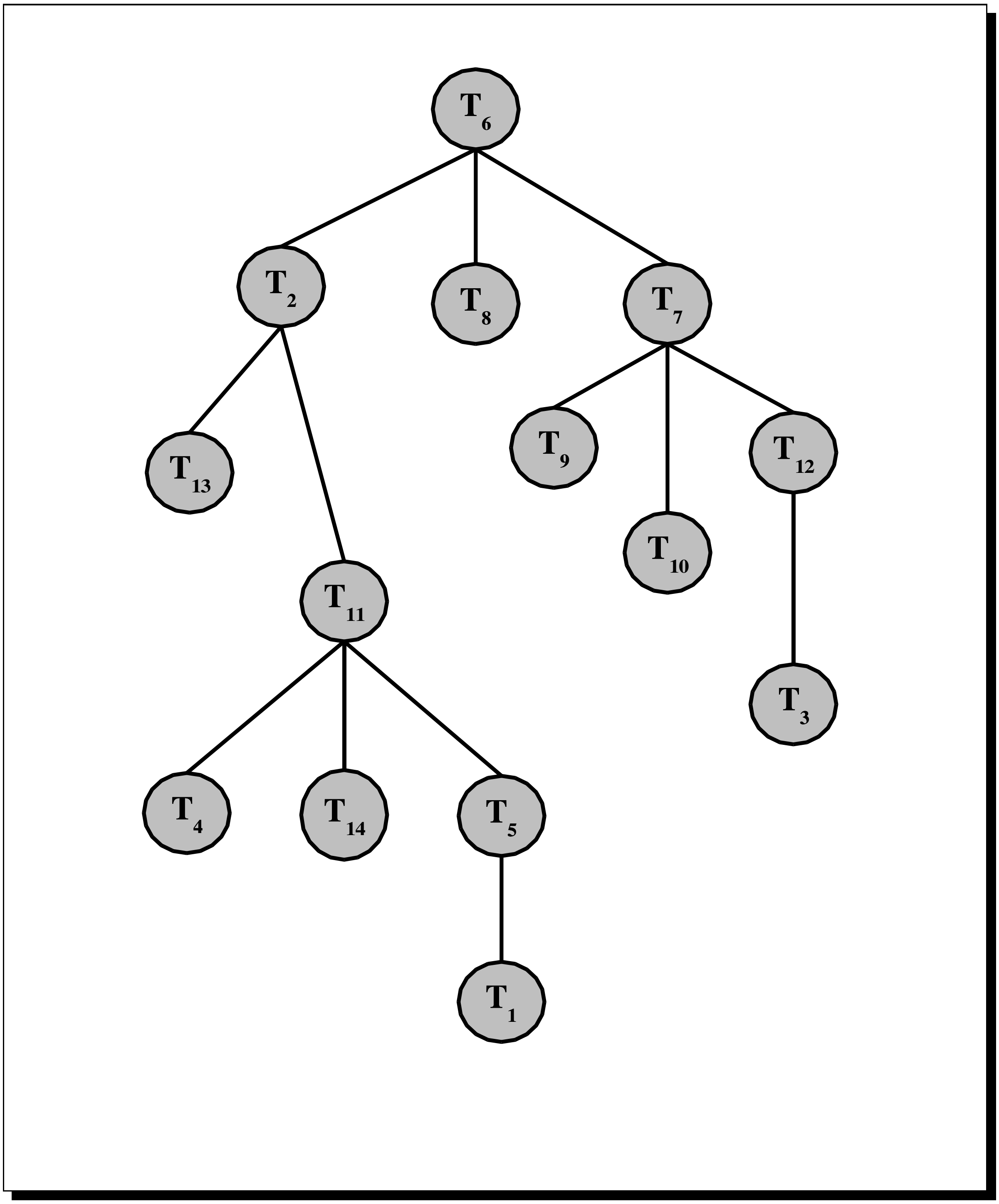


Figura 2.1.: Árbol generador de peso máximo construido tras aplicar el paso quinto del Algoritmo 2.1.

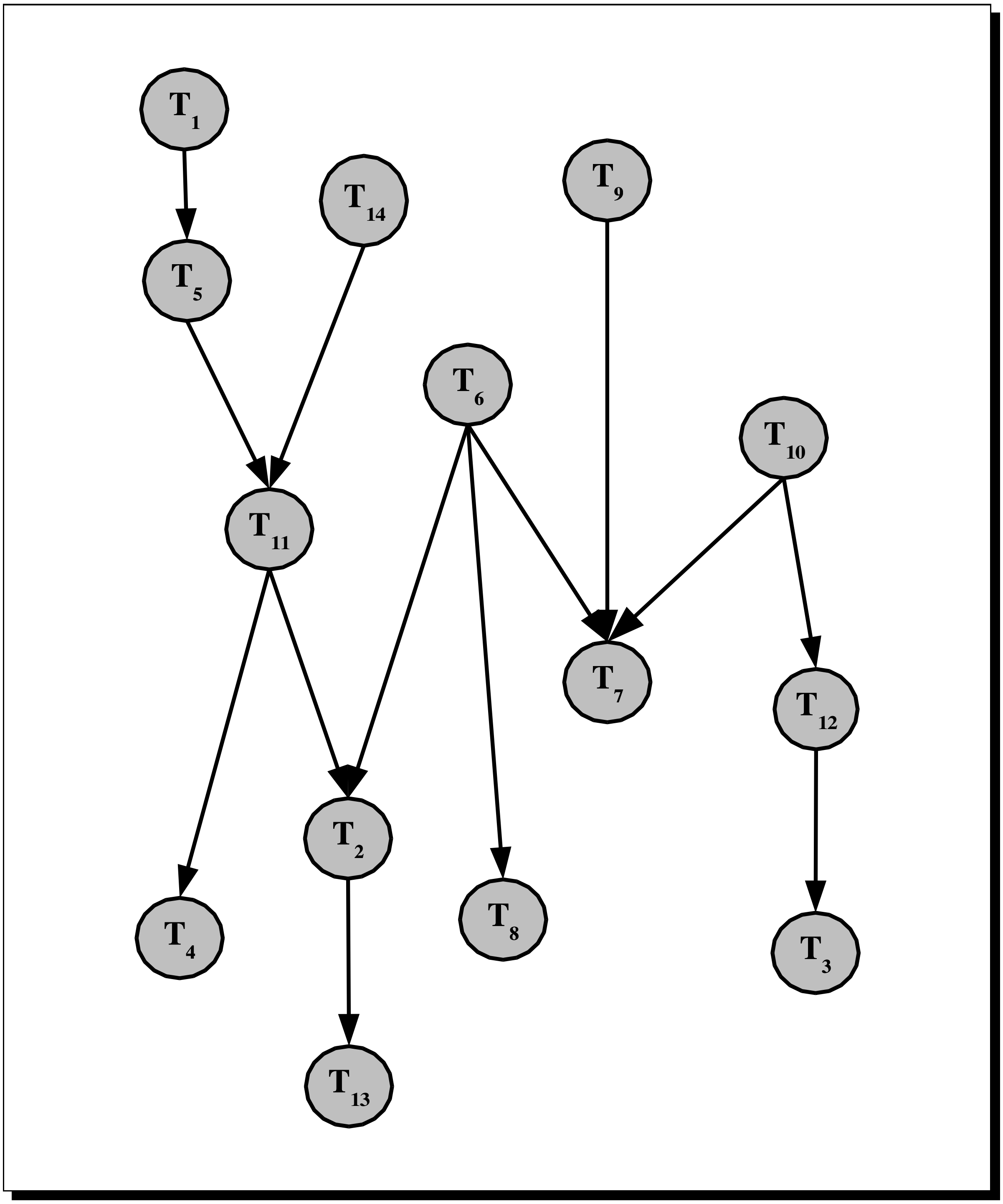


Figura 2.2.: Poliárbol aprendido tras ejecutar la fase de orientación de enlaces.

- El proceso de creación del árbol generador de peso máximo se mejora incluyendo un test de independencia para asegurar que las aristas que se introducen unan términos realmente dependientes.
- En cuanto a la estructura, se ha aprendido un poliárbol en lugar de un árbol como se hace en [Rij77, Rij79, HR78, RHP81, FC89, Fri88, SD91, Sav92], ganando así precisión ya que se aumenta la capacidad de representación de dependencias e independencias con respecto al modelo real.

2.3.2. La estimación de las distribuciones de probabilidad almacenadas en la red.

Por último, y una vez que la estructura del poliárbol ha sido construida, el algoritmo tiene que calcular (a partir del fichero invertido) las probabilidades a priori para los nodos raíces, y las probabilidades condicionales del resto de nodos dados todos sus padres, con lo que estarán las relaciones cuantitativas entre los nodos de la red totalmente especificadas, y por tanto, la red bayesiana quedará completa y lista para ser usada en el proceso de expansión.

Aunque en el capítulo siguiente se hará un estudio más detallado sobre este aspecto, en éste vamos a introducir los métodos básicos que se han diseñado para estimar estas probabilidades, y con los que se ha realizado la experimentación sobre expansión de consultas:

- Distribuciones de probabilidad en los nodos raíces:

  Dado un nodo raíz que representa la variable binaria aleatoria T_i, éste deberá almacenar la distribución de probabilidad marginal, es decir, la probabilidad de que T_i no sea relevante dado el vacío y la probabilidad de que T_i sea relevante dado el vacío. Por tanto, cada nodo T_i que no tenga padres alojará su correspondiente probabilidad a priori:

  $$\begin{array}{rcl} p(T_i \; no \; relevante) & = & p(\bar{t}_i) \\ p(T_i \; relevante) & = & p(t_i) \end{array}$$

  El estimador diseñado calcular la probabilidad a priori es el siguiente:

  - *pp2*: determina la probabilidad de relevancia de un término como nuestra creencia a priori sobre el hecho de que el término T_i puede pertenecer a una consulta. En este caso, suponemos que todos los términos de la colección tienen la misma probabilidad de pertenecer a la consulta, por lo que le asociamos la siguiente probabilidad:

    $$p(t_i) = \frac{1}{M} \text{ y } p(\bar{t}_i) = 1 - p(t_i),$$

    donde M es el número de términos pertenecientes a la colección.

- Distribuciones de probabilidad en el resto de nodos:

  Seguidamente nos vamos a centrar en el método diseñado para estimar las distribuciones de probabilidad condicionada de los nodos que tienen padres en el grafo. Estas distribuciones representan $P(T_i \mid T_1, \ldots, T_p)$, es decir, los valores de probabilidad condicionada de todas las combinaciones posibles de valores que toman los nodos padres de un nodo, T_i, que notaremos por $\Pi(T_i)$, para cada uno de los dos valores que toma la variable representando al término T_i. Una configuración de valores dados para $\Pi(T_i)$ se representará por $\pi(T_i)$.

  Previamente, presentaremos la notación usada. Dado un conjunto de términos $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$, 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 $T_i \in \mathcal{T}$. Así, $\mathbf{t_i} = t_i$ si la variable i-ésima de $\mathcal{T}$ es relevante y $\mathbf{t_i} = \bar{t}_i$, en caso de que no lo sea. Por ejemplo, para $\mathcal{T} = \{T_1, T_2, T_3, T_4\}$, 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 $\mathcal{T}$ 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.

  El estimador diseñado lo denominaremos *pc-mv*, corresponde con un estimador de máxima verosimilitud [Goo65] y tiene la siguiente forma:

  $$p(t_i \mid \pi(T_i)) = \frac{n(< t_i, \pi(T_i) >)}{n(\pi(T_i))} \text{ y } p(\bar{t}_i \mid \pi(T_i)) = 1 - p(t_i \mid \pi(T_i))$$

  Este estimador será uno de los utilizados en el siguiente capítulo cuando se use el poliárbol de términos como uno de los componentes básicos del modelo de recuperación que en él presentamos.

2.4. Aplicación del tesauro a la expansión de consultas.

Una vez que el tesauro está listo para su uso, para determinar su calidad y funcionalidad procederemos a emplearlo como ayuda para llevar a cabo un proceso de expansión de consultas, el cual queda descrito en el algoritmo 2.2. En este proceso intervienen dos elementos claramente definidos:

- Un S.R.I. cualquiera, que será el sistema en el que el usuario efectúe su consulta y el que realice la recuperación de documentos, además de llevar a cabo la evaluación de la consulta (en nuestro caso, se ha utilizado el S.R.I. SMART).

- Un subsistema de expansión de consultas, compuesto por el tesauro y un mecanismo de inferencia, mediante el cual se obtendrán los nuevos términos que se añaden a la consulta original (la red bayesiana que hace las veces de tesauro y el mecanismo de propagación de probabilidades en poliárboles).

Algoritmo 2.2 Expansión de consultas.

entradas: tesauro de una colección (poliárbol de términos) y un conjunto de consultas.
salidas: para cada consulta, un conjunto de términos con los que expandirla.
para cada consulta, Q, suministrada al subsistema de expansión por el S.R.I. principal **hacer**
  para cada término T_i de la consulta **hacer**
    Buscar el término en el grafo.
    Asignarle $\lambda(T_i) = (0,1)$. {Instanciarlo a relevante}
  fin para
  Ejecutar el proceso de propagación de probabilidades en el poliárbol y calcular para cada T_i su probabilidad a posteriori, es decir, $p(t_i \mid Q)$.
  $CE \leftarrow \emptyset$
  para cada término, T_i, del poliárbol **hacer**
    si $p(t_i \mid Q) > umbral$ **entonces**
      $CE \leftarrow CE \cup \{(T_i, p(t_i \mid Q)\}$
    fin si
  fin para
  Devolver al S.R.I. el conjunto de términos de expansión CE para la consulta Q.
fin para

Dada una consulta formulada a un S.R.I., el proceso de expansión de consultas comienza situando las evidencias en el árbol aprendido. Esta acción consiste en buscar los términos que aparecen en la consulta en el poliárbol y asignar sus estados a "el término es relevante". Después de eso, se efectúa un proceso de propagación. Debido a que nuestra red es un poliárbol, podemos usar un método de inferencia exacto y eficiente para propagar las probabilidades [Pea88]. Como resultado de la propagación obtenemos la probabilidad de que un término sea relevante dado que todos los términos de la consulta son relevantes, para cada nodo. El siguiente paso es obtener aquellos términos con mayor probabilidad. Para seleccionar los términos que añadiremos a la consulta, usamos un valor umbral que establece un límite inferior, es decir, aquellos términos cuya probabilidad a posteriori sea mayor que un umbral dado se toman para ser añadidos a la consulta, formando así nuestra consulta expandida. Los pesos tf de los términos expandidos son precisamente sus correspondientes probabilidades a posteriori. Una vez que el nuevo vector de la consulta esté completamente creado, SMART toma esa nueva consulta y efectúa una recuperación, obteniendo el conjunto de documentos relevantes.

Este enfoque tiene una clara ventaja con respecto a otros existentes en la literatura y es que la expansión se hace de manera global a la consulta, y no individual a cada término. Así, los términos que se añaden a la consulta son los que alcanzan una probabilidad a posteriori más alta atendiendo a la consulta completa.

2.5. Experimentación.

Para determinar la validez del tesauro presentado en este capítulo, hemos experimentado con las colecciones ADI, CRANFIELD y MEDLARS. Se han elegido estas tres debido a que, dentro de las colecciones de prueba estándar normalmente utilizadas, tienen tamaños pequeño, medio y grande, respectivamente, pudiendo determinar así el impacto del número de términos con respecto a la calidad del poliárbol aprendido, y posteriormente, al rendimiento de la expansión.

Una vez que se ha expandido la consulta, ésta se pasará a SMART para que recupere. Para medir la calidad de la recuperación con la expansión, se calculará la curva Exhaustividad - Precisión para los once puntos estándar de exhaustividad, y se comparará con la conseguida por SMART con la consulta original, calculando posteriormente los porcentajes de cambio de la primera con respecto a la segunda. Comentar, por último, que los pesos asignados a los documentos y a las consultas empleados por SMART se corresponden al esquema *nnn*, o lo que es lo mismo, la frecuencia de aparición del término en el documento o la consulta (el peso *tf*).

El diseño de experimentos que hemos llevado a cabo ha tenido en cuenta los parámetros existentes en los dos procesos básicos implicados en la expansión: el aprendizaje y la propia expansión. Éstos son los siguientes:

- En el aprendizaje: nivel de confianza para los dos test de independencia realizados.
- En la selección: la red bayesiana aprendida y el umbral con el que seleccionarán los términos.

Así, la batería de experimentos diseñada para cada colección está compuesta por:

- Niveles de confianza: 90 %, 95 %, 97.5 %, 99 % y 99.5 %.
- Umbrales de selección: 0.5, 0.6, 0.7, 0.8 y 0.9.

Teniendo en cuenta estos argumentos, para cada colección se han aprendido un total de cinco poliárboles, cada uno con un nivel de confianza. El siguiente paso ha sido expandir la batería de consultas que trae cada colección con cada uno de los cinco tesauros, ejecutando la expansión cinco veces, una por umbral, alcanzando de esta manera un total de veinticinco experimentos.

Las tablas 2.1, 2.2 y 2.3 muestran algunos de los resultados obtenidos para los experimentos previamente comentados para las colecciones ADI, CRANFIELD y MEDLARS. En cada tabla se ofrecen la curva de exhaustividad-precisión y la precisión media de todos los valores de exhaustividad (fila etiquetada como *Media*), utilizando SMART de manera autónoma (columna *Precisión SMART*) y SMART con expansión de consultas (columna *Precisión SMART + E.C.*). También se muestra el porcentaje de cambio con respecto a los resultados obtenidos por SMART con las consultas originales (%C). El valor de la precisión que se muestra para cada nivel de exhaustividad fijo es la media para todas las consultas de la colección. Las últimas dos filas en cada tabla representan la exhaustividad y la precisión obtenidas de la recuperación de

Exhaustividad	Precisión SMART	Precisión SMART + E.C.	%C	Media 25 exp.
0.0	0.4824	0.5052	4.73	0.4992
0.1	0.4824	0.5052	4.73	0.4992
0.2	0.4343	0.4616	6.29	0.4642
0.3	0.4203	0.4365	3.85	0.4403
0.4	0.3640	0.3987	9.53	0.3916
0.5	0.3405	0.3878	13.89	0.3856
0.6	0.2775	0.3392	22.23	0.3053
0.7	0.2247	0.2761	22.87	0.2490
0.8	0.2143	0.2653	23.80	0.2423
0.9	0.1874	0.2417	28.98	0.2123
1.0	0.1863	0.2405	29.09	0.2111
Media	0.3285	0.3689	15.45	0.3545
Exhaustividad	0.5036	0.5983	18.80	0.5746
Precisión	0.1524	0.1695	11.22	0.1679

Cuadro 2.1.: Resultados para la expansión con ADI (Nivel de confianza= 95 %; umbral= 0.5).

un número fijo de documentos (nosotros utilizamos el valor por defecto de SMART, establecido en 15). Para cada colección, mostramos los resultados del uso de la expansión de consultas sólo para un experimento, así como la media de los resultados de los veinticinco experimentos.

Para las colecciones ADI y MEDLARS se han obtenido mejores resultados en los veinticinco experimentos que los ofrecidos con las consultas originales por SMART, en términos tanto de precisión y exhaustividad, como en número de documentos recuperados. Así, nuestro método de expansión de consultas parece bastante robusto con respecto a estas dos colecciones. Los mejores valores medios de precisión se alcanzan más a menudo en umbrales bajos - medios (0.5 - 0.7). Los valores de precisión en nuestros experimentos para los diez puntos de exhaustividad son siempre mejores que aquellos pertenecientes al conjunto original de consultas, estableciendo diferencias mayores en los valores de exhaustividad medios - altos. En estas dos colecciones no hay diferencias importantes entre las cinco redes, correspondientes a los cinco niveles de confianza usados en los tests de independencia.

Para la colección CRANFIELD, se puede observar cómo el umbral utilizado para seleccionar los términos con los que expandir las consultas es un parámetro importante en el rendimiento de la recuperación: cuanto más alto es, mejores resultados se obtienen. Las expansiones no son buenas del todo para valores bajos del umbral, aunque mejoran cuando este límite es más alto: sólo obtenemos mejores resultados que los ofrecidos con las consultas originales utilizando umbrales altos (0.8 y 0.9) (esto explica las medias bastante pobres mostradas en la última columna de la tabla 2.2). Por el contrario, no hay diferencias apreciables entre los resultados de las diferentes redes, comportándose de forma muy similar. Estos resultados sugieren que en este caso, la expansión de consultas es útil cuando el número de términos añadido a la consulta es bastante pequeño (hecho que probablemente sea debido a las características de esta colección).

Exhaustividad	Precisión SMART	Precisión SMART + E.C.	%C	Media 25 exp.
0.0	0.5730	0.5606	-2.16	0.5345
0.1	0.5236	0.5150	-1.64	0.4887
0.2	0.4158	0.4177	0.46	0.3926
0.3	0.3042	0.3105	2.07	0.2962
0.4	0.2452	0.2558	4.32	0.2415
0.5	0.2171	0.2289	5.44	0.2124
0.6	0.1687	0.1810	7.29	0.1675
0.7	0.1183	0.1263	6.76	0.1187
0.8	0.0931	0.1016	9.13	0.0927
0.9	0.0656	0.0696	6.10	0.0650
1.0	0.0573	0.0609	6.28	0.0569
Media	0.2529	0.2570	4.00	0.2424
Exhaustividad	0.3397	0.3450	1.56	0.3284
Precisión	0.1591	0.1636	2.83	0.1576

Cuadro 2.2.: Resultados para la expansión con CRANFIELD (Nivel de confianza= 97.5; umbral= 0.9).

No sabemos exactamente por qué nuestro método tiene un mayor rendimiento con las colecciones ADI y MEDLARS que con CRANFIELD. Contestar esta pregunta requerirá un estudio en profundidad de las características de estas colecciones (especificidad y generalidad de las consultas, términos de indexación empleados, número de términos por consulta, etc.).

En general, podemos concluir que la mejora en la efectividad de la recuperación inducida por nuestro método es importante, pero, a la vez, moderada. Este hecho nos anima a continuar estudiando este campo, con objeto de obtener mejoras más sensibles. Otras conclusiones son que el nivel de confianza de los tests de independencia no es relevante para el rendimiento de la consulta expandida, y por otro lado, que el umbral impuesto para la selección de los términos a añadir a la consulta original puede ser bastante relevante. Esto sugiere la necesidad de desarrollar algunas heurísticas para seleccionar automáticamente el umbral de acuerdo a los datos de una colección dada (por ejemplo, tener en cuenta las frecuencias inversas de los términos en las consultas). De cualquier forma, nos parece que la mejora del rendimiento de la recuperación dependerá de la colección.

Exhaustividad	Precisión SMART	Precisión SMART + E.C.	%C	Media 25 exp.
0.0	0.7235	0.7324	1.23	0.7345
0.1	0.6389	0.7042	10.22	0.6788
0.2	0.5810	0.6110	5.16	0.6016
0.3	0.5204	0.5598	7.57	0.5505
0.4	0.4561	0.4938	8.27	0.4865
0.5	0.3725	0.3996	7.28	0.4013
0.6	0.2887	0.3381	17.11	0.3297
0.7	0.2403	0.2894	20.43	0.2786
0.8	0.1956	0.2367	21.01	0.2292
0.9	0.1534	0.1752	14.21	0.1725
1.0	0.0875	0.1115	27.43	0.1033
Media	0.3871	0.4229	12.72	0.4151
Exhaustividad	0.3161	0.3406	7.75	0.3356
Precisión	0.4467	0.4800	7.45	0.4738

Cuadro 2.3.: Resultados para la expansión con MEDLARS (Nivel de confianza= 97.5; umbral= 0.7).

3. Desarrollo de modelos de recuperación de información basados en redes bayesianas.

En este capítulo vamos a presentar el modelo de recuperación basado en redes bayesianas que hemos desarrollado. En él especificaremos la topología del mismo, los métodos para estimar las matrices de probabilidades que almacena y la manera de realizar la inferencia, estableciendo las diferencias y similitudes con los tres principales modelos ya expuestos en el capítulo 1. Además, y como forma de medir la capacidad recuperadora del modelo y, por tanto, su calidad, expondremos los resultados de la experimentación efectuada sobre las colecciones ADI, CACM, CISI, CRANFIELD y MEDLARS, demostrando así su aceptable comportamiento recuperador, totalmente comparable a los otros modelos.

3.1. Introducción.

En el capítulo anterior utilizamos las redes bayesianas para realizar un proceso de expansión de consultas. El poliárbol de términos sólo ayudaba en una tarea complementaria a la recuperación, la expansión de la consulta, siendo un S.R.I. externo, concretamente SMART, quien efectuaba la recuperación propiamente dicha una vez que la consulta original había sido modificada. Esta expansión se llevaba a cabo mediante la selección de aquellos términos representados en el poliárbol de términos que se habían considerado más probables tras efectuar la propagación de una consulta. En este capítulo vamos a presentar un sistema de recuperación de información basado completamente en redes bayesianas. Para representar el conocimiento usando redes bayesianas, la tarea inicial consiste en seleccionar aquellas variables relevantes al problema que vamos a modelar. En el ámbito de la R.I. podemos encontrar dos conjuntos de variables claramente diferenciadas:

- *Variables término*, que se corresponden con los términos pertenecientes al glosario de la colección, que notaremos como $\mathcal{T}$. Este tipo de variables ya fue descrito en el capítulo anterior, aunque a modo de recordatorio diremos que las representaremos 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: $\{\bar{t}_i, t_i\}$. Semánticamente, $\bar{t}_i$ equivale a *"el término T_i no es relevante"* y t_i se corresponde con *"el término T_i es relevante"*.

- *Variables documento*, que representan a cada uno de los documentos que componen el fondo bibliográfico, que denominaremos $\mathcal{D}$. Estas variables se notarán como D_j, para $j = 1,\ldots,N$, con N el número total de documentos. Toman los valores del conjunto $\{\bar{d}_j, d_j\}$, donde $\bar{d}_j$ representa al suceso *"el documento D_j no es relevante"* y d_j hace lo propio con *"el documento D_j es relevante"*.

Una vez seleccionadas las variables relevantes a nuestro problema, el siguiente paso a la hora de diseñar el S.R.I. basado en redes bayesianas consiste en determinar qué relaciones existen entre ellas, o lo que es lo mismo, determinar la estructura de la red. Para ello, podemos considerar tres enfoques distintos:

1. Utilizar la información suministrada por un experto documentalista para construir el modelo, permitiéndonos conocer las relaciones de relevancia existentes entre los términos y los documentos. Esta aproximación es prácticamente inviable, debido tanto al volumen de datos existente, como a la alta velocidad con que nuevos documentos suelen llegar al S.R.I.

2. Aplicar un algoritmo de aprendizaje para obtener la estructura de la red. Este algoritmo será capaz de construir un modelo a partir de la información proporcionada por el conjunto de documentos que componen la colección. Esta solución también se muestra poco factible debido al gran número de variables involucradas (del orden de cientos de miles en una colección media).

3. Aplicar un método híbrido, que recoja las ventajas de ambos enfoques anteriores.

Este último enfoque será el que adoptaremos para desarrollar nuestro modelo. Así, por un lado, nos propondremos utilizar, como "conocimiento experto", un conjunto de principios de coherencia que debe preservar nuestro modelo y que servirán de guía a la hora de construirlo, y por otro, nos propondremos dotar al modelo de la capacidad de inferir determinadas relaciones entre las variables implicadas a partir de la colección documental. En este caso, el conjunto de suposiciones que estableceremos será el siguiente:

1. Existe una relación fuerte entre un documento y cada uno de los términos que lo indexan. Como consecuencia, para cada término que indexe un documento debe existir un enlace entre el nodo que corresponde a ese término y el nodo asociado al documento que indexa.

2. Las relaciones entre documentos sólo se dan a través de los términos que contienen dichos documentos. Esta suposición implica que no existan enlaces que conecten los nodos documento entre sí. Si consideramos que, para un S.R.I., un documento no es más que la lista de términos que lo indexan (la representación que maneja dicho S.R.I.), entonces parece lógico además que las interacciones entre documentos sólo puedan hacerse por medio de términos comunes.

3. 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 o término T_k sea relevante o irrelevante. Esta suposición implica que, probabilísticamente, los documentos sean condicionalmente independientes dados los términos por los que han sido indexados. Gráficamente, los enlaces que unen los términos y los documentos en el grafo quedarán dirigidos desde los términos hacia los documentos. Esta hipótesis parece también razonable, teniendo en cuenta de nuevo el hecho de que, para un S.R.I., un documento no tiene más entidad que la que le proporciona la suma de sus términos. La alternativa de dirigir los enlaces desde los documentos a los términos implicaría que los documentos fuesen marginalmente independientes, es decir, que afirmar que un documento es relevante (para alguna consulta) no aporta información alguna sobre la posible relevancia de otros documentos. Esta conclusión está en franca contradicción con algunos principios claramente establecidos en R.I., como por ejemplo, la *hipótesis del agrupamiento* [Rij79].

Estas dos últimas suposiciones se relajarán en el capítulo 5 de tal forma que se puedan establecer relaciones directas entre documentos, con objeto de enriquecer la expresividad del modelo.

Teniendo en cuenta estos tres principios anteriores, en nuestro modelo podemos distinguir entre dos capas de nodos claramente diferenciadas (figura 3.1), la capa de nodos término y la de nodos documento, que dan lugar a dos subgrafos distintos pero enlazados entre sí, que denominaremos *subred de términos* y *subred de documentos*, respectivamente. Los arcos irán dirigidos desde los nodos de la primera capa a los de la segunda, respetando así las relaciones de independencia condicional antes mencionadas. Otra razón que apoya esta decisión es que parece más intuitivo hablar de probabilidad de relevancia de documentos dados los términos que al contrario.

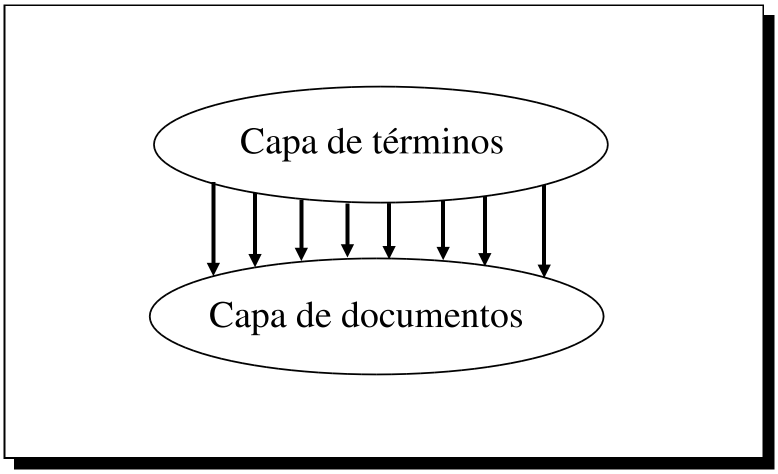


Figura 3.1.: División de las variables de la red bayesiana documental en capas de nodos término y documento.

Tomando como base este esqueleto genérico, en este capítulo desarrollamos tres modelos distintos, que variarán dependiendo de las estructuras que se utilicen para representar las relaciones entre los términos de la colección.

Una vez que la red bayesiana documental se haya construido, ya está lista para ser usada por el S.R.I. Así, dada una consulta efectuada a la aplicación, el proceso de recuperación comienza situando las evidencias en la red, es decir, asignando a las variables asociadas con los términos que aparecen en la consulta el valor "el término es relevante". En ese momento está todo listo para realizar la propagación de probabilidades y obtener la probabilidad a posteriori de cada documento, que representará la probabilidad de que un documento sea relevante dado que los términos de la consulta son también relevantes. Finalizados los cálculos, el último paso que debe dar el S.R.I. es el de ordenar los documentos por dicha probabilidad a posteriori y, por un lado, ofrecerle al usuario dicha lista para que la inspeccione y, por otro, medir el rendimiento del sistema para determinar la calidad recuperadora del mismo.

Pasamos a detallar brevemente el contenido de este capítulo. Si se consideran todos los términos independientes entre sí, obtendremos un grafo en donde no existe ningún enlace conectando los términos, dando lugar al modelo *subred de términos simple*, que presentaremos en la siguiente sección. Por el contrario, si se descarta la independencia entre términos, necesitaremos de un mecanismo para representar el conjunto de relaciones de dependencia. Dependiendo de la topología de la estructura resultante obtendremos dos modelos distintos: la *subred de términos aumentada*, estudiada en la sección 3.3, y la *subred de términos mixta*, que expondremos en la sección 3.4. La siguiente sección, la 3.5, se centrará en cómo se realiza la inferencia en estos modelos, con objeto de recuperar documentos. La calidad de los tres modelos propuestos ha sido medida mediante el desarrollo de un conjunto de baterías de experimentos, cuyos resultados y comentarios los presentaremos en la sección (3.6). Para concluir, en la sección 3.7 ofrecemos una comparativa con el resto de modelos de recuperación basados en redes bayesianas.

Por abuso del lenguaje, y teniendo en cuenta que los documentos están aislados entre sí independientemente de la topología de la subred de términos, cuando hablemos de red simple, aumentada o mixta, nos estaremos refiriendo a la red bayesiana documental con la topología simple, aumentada o mixta de la subred de términos.

3.2. La red simple.

El modelo que presentamos en esta sección se va a denominar *red bayesiana documental basada en la subred de términos simple*, o lo que es lo mismo, *la red simple*. En este modelo, se considera una suposición adicional: los términos son independientes entre sí. Como consecuencia, su topología está compuesta por dos subredes, dentro de las cuales los nodos están aislados unos de otros. Los únicos enlaces existentes serán los que conectan ambas redes mediante arcos que nacen de los nodos término y apuntan a los nodos documento. Cada nodo documento recibirá un arco proveniente de cada uno de los términos con los que ha sido indexado.

Esta topología no necesita de un algoritmo de aprendizaje para su construcción, ya que, dada una colección cualquiera $\mathcal{D}$, la confección de la red simple se hace de manera inmediata tal y como se ha explicado en el párrafo anterior.

En la figura 3.2 se observa gráficamente la red simple obtenida a partir de la colección siguiente, la cual está compuesta por cuatro documentos y un total de doce términos diferentes:

$$\begin{aligned} D_1 &= (T_1, T_2, T_3, T_4) \\ D_2 &= (T_4, T_5, T_6, T_7) \\ D_3 &= (T_4, T_6, T_8, T_9, T_{10}) \\ D_4 &= (T_8, T_{10}, T_{11}, T_{12}) \end{aligned}$$

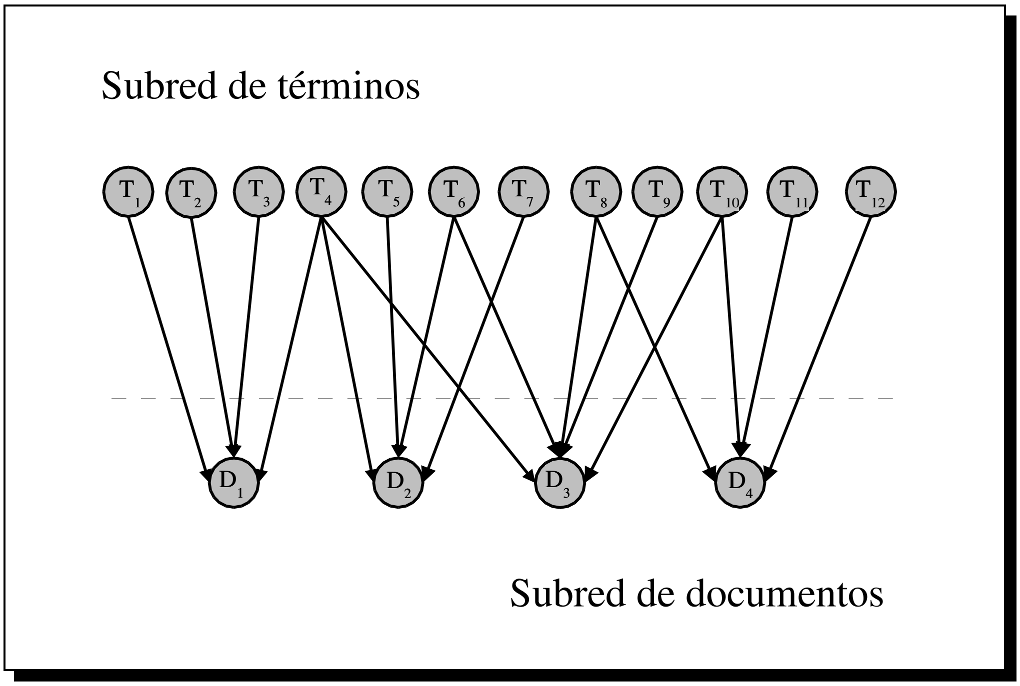


Figura 3.2.: Red bayesiana documental formada por la subred de términos simple.

3.2.1. Estimación de la información cuantitativa.

Una vez creada la estructura de la red simple, para tener totalmente especificada la red bayesiana, nos quedaría estimar la fuerza de las relaciones representadas, lo que implica realizar la estimación de las distribuciones de probabilidad de cada uno de los nodos de la red. Todos los nodos término, por ser raíces, almacenarán las correspondientes distribuciones de probabilidad marginales, mientras que para los nodos documento se almacena un conjunto de distribuciones de probabilidad condicionadas.

En los dos siguientes apartados trataremos cómo se calculan estas dos distribuciones de probabilidad y la manera en que se resuelven algunos problemas planteados en su estimación y en su posterior uso por parte de los mecanismos de inferencia.

3.2.1.1. Estimación de las distribuciones de probabilidad marginales.

Como ya hemos comentado, todos los nodos que sean raíces deberán contener su distribución marginal. Así, dado un nodo raíz que representa a la variable T_i, éste deberá almacenar la probabilidad de que T_i no sea relevante y la probabilidad de que T_i sea relevante, que notaremos mediante:

$$\begin{aligned} p(\bar{t}_i) &= p(T_i \text{ no relevante}) \\ p(t_i) &= p(T_i \text{ relevante}) \end{aligned}$$

Centrándonos en la manera de estimar esas probabilidades, podemos clasificar los métodos estudiados en dos grupos:

1. *Estimadores sin información*: no hacen distinción entre términos, asignando las mismas probabilidades a cada uno de los términos. Se han diseñado dos estimadores dentro de este grupo:

   - *pp1*: supone que a priori no se tiene el conocimiento suficiente para discernir si el término es relevante o no; por tanto, le deberemos asociar al término el estado de mayor incertidumbre, el cual se representa en términos probabilísticos como una distribución uniforme:

     $$p(t_i) = 0,5 \text{ y } p(\bar{t}_i) = 0,5$$

   - *pp2*: El estimador ya introducido en el capítulo anterior y cuya expresión, a modo de recordatorio, corresponde con

     $$p(t_i) = \frac{1}{M} \text{ y } p(\bar{t}_i) = 1 - p(t_i),$$

     donde M es el número de términos pertenecientes a la colección.

2. *Estimadores con información*: las probabilidades calculadas dependen de la calidad del término, medida sobre la base del número de documentos en los que aparece en la colección. Los estimadores diseñados son:

   - *pp3*: considera la frecuencia con que un término T_i aparece en la colección, n_i, indicativa de la frecuencia con que se espera que aparezca en el conjunto de consultas. Por tanto, podemos calcular la probabilidad a priori como:

$$p(t_i) = \frac{n_i}{N} \text{ y } p(\bar{t}_i) = 1 - p(t_i)$$

El problema que podría presentar esta aproximación es que se considera que a priori son más relevantes aquellos términos que aparecen más en la colección, esto es, los más frecuentes. Sin embargo, suele ocurrir que los términos más específicos, con menor peso, son los que para una determinada consulta permiten discernir mejor si un documento es relevante o no. Para tener en cuenta este hecho, podemos utilizar la siguiente expresión (*pp4*), donde se penaliza a los términos que ocurren más frecuentemente en la colección:

- *pp4*: este método obtiene la probabilidad de relevancia de un término como nuestra creencia a priori de que éste pertenezca a la consulta, medida utilizando la expresión *pp2*, ponderada por la especificidad del término, que determina la capacidad de discriminar los documentos que lo contienen. Así, un término que aparezca en pocos documentos tendrá una mayor especificidad que otro que sea muy frecuente en la colección. Dicha especificidad queda calculada como $(N - n_i)/N$. Por tanto, las probabilidades almacenadas en el nodo son:

$$p(t_i) = \frac{N - n_i}{N} \frac{1}{M} \text{ y } p(\bar{t}_i) = 1 - p(t_i)$$

3.2.1.2. Estimación de las distribuciones de probabilidad condicionadas.

El último paso que se debe dar antes de poder utilizar para la recuperación este modelo, es la estimación de las distribuciones de probabilidad almacenadas en los nodos documento, es decir, $p(D_j \mid T_1, \ldots, T_{m_j})$.

El principal problema que plantea la estimación de estas matrices consiste en que el tamaño de éstas será exponencial en el número de términos que indexan el documento, o lo que es lo mismo, y hablando en términos de la estructura, exponencial en el número de padres de un nodo documento. Así, si un documento ha sido indexado por m_j términos, y teniendo en cuenta que cada término representa una variable aleatoria binaria, el número de distribuciones de probabilidad a estimar será 2^{m_j}. Por ejemplo, el documento D_1 de la figura 3.2 almacenará 2^4 distribuciones de probabilidad condicionadas (una para cada posible configuración de los padres) [1]. En concreto:

$$\begin{array}{cccc}
p(d_1 \mid \bar{t}_1,\bar{t}_2,\bar{t}_3,\bar{t}_4) & p(d_1 \mid \bar{t}_1,\bar{t}_2,\bar{t}_3,t_4) & p(d_1 \mid \bar{t}_1,\bar{t}_2,t_3,\bar{t}_4) & p(d_1 \mid \bar{t}_1,\bar{t}_2,t_3,t_4) \\
p(d_1 \mid \bar{t}_1,t_2,\bar{t}_3,\bar{t}_4) & p(d_1 \mid \bar{t}_1,t_2,\bar{t}_3,t_4) & p(d_1 \mid \bar{t}_1,t_2,t_3,\bar{t}_4) & p(d_1 \mid \bar{t}_1,t_2,t_3,t_4) \\
p(d_1 \mid t_1,\bar{t}_2,\bar{t}_3,\bar{t}_4) & p(d_1 \mid t_1,\bar{t}_2,\bar{t}_3,t_4) & p(d_1 \mid t_1,\bar{t}_2,t_3,\bar{t}_4) & p(d_1 \mid t_1,\bar{t}_2,t_3,t_4) \\
p(d_1 \mid t_1,t_2,\bar{t}_3,\bar{t}_4) & p(d_1 \mid t_1,t_2,\bar{t}_3,t_4) & p(d_1 \mid t_1,t_2,t_3,\bar{t}_4) & p(d_1 \mid t_1,t_2,t_3,t_4)
\end{array}$$

[1]Los valores de no relevancia del documento se obtienen por dualidad.

Teniendo en cuenta que en una colección de tamaño normal, el número de términos que indexan un documento puede ser de 100 ó 200, el total de combinaciones posibles es enorme, originando la siguiente sucesión de problemas:

- El tiempo necesario para estimar las probabilidades condicionadas de cada nodo puede ser extremadamente largo.
- Si se emplean datos para estimar esas distribuciones, la fiabilidad de tales estimaciones será muy pequeña, salvo el (improbable) caso de disponer de una cantidad de datos gigantesca.
- En caso de que se hayan podido calcular las matrices, el espacio en disco requerido para su almacenamiento sería gigantesco.
- Y finalmente, si ha sido posible guardarlas, los métodos de propagación se harán muy lentos debido a la gran cantidad de tiempo que necesitarán para poder gestionar las matrices almacenadas en la red.

La existencia de estos cuatro 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*. Básicamente, una función de este tipo se puede considerar como una representación implícita de una distribución de probabilidad sobre el conjunto de variables $x_1, x_2, \ldots, x_n$. Genéricamente, una función de probabilidad tiene la forma $fp(<\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}>)$, donde $<\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}>$ representa una configuración para $x_1, x_2, \ldots, x_n$, siendo $\mathbf{x_i}$ uno de los posibles valores que puede tomar la variable x_i. La función de probabilidad devolverá el valor de probabilidad asociado a esta configuración.

En nuestro caso particular, las variables involucradas serán un documento D_j y el conjunto de términos por el que ha sido indexado: $\Pi(D_j) = \{T_1, T_2, \ldots T_{m_j}\}$, de forma que $fp(\mathbf{d_j}, <\mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_{m_j}}>)$ devuelve la probabilidad condicionada $p(\mathbf{d_j} \mid \mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_{m_j}})$, donde $<\mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_{m_j}}>$ es una configuración particular para $\Pi(D_j)$ y $\mathbf{d_j} = d_j$ si el documento es considerado relevante o $\mathbf{d}_j = \bar{d}_j$ si no lo es. Cada $\mathbf{t_i}$ en $\pi(D_j)$ es el valor que toman las variables en la configuración, esto es, $\mathbf{t_i} = t_i$ si el término es relevante y $\mathbf{t_i} = \bar{t}_i$ si no es relevante.

En la etapa de inferencia, las funciones de probabilidad calcularán la probabilidad condicionada requerida en el momento en que se necesite. De esta manera, la representación explícita de la distribución de probabilidad condicional se sustituye por una implícita, representada por la función usada. Si durante la ejecución del proceso de propagación, éste necesita la probabilidad concreta, por ejemplo, $p(d_1 \mid \bar{t}_1, t_2, \bar{t}_3, t_4)$, se invocará la función de probabilidad y ésta calculará en ese momento el valor buscado.

Una vez resuelto el problema de la representación de las distribuciones consideradas, pasamos a estudiar cómo podemos calcular el valor concreto que devolverá la función de probabilidad. El cálculo de este valor se basará en otras probabilidades más simples y más fáciles de estimar, las cuales serán combinadas para obtener el valor de la probabilidad condicionada requerida.

Supongamos un documento D_j indexado por los términos $T_1,\ldots,T_{m_j}$. El cálculo de la función de probabilidad se hará en dos pasos: en el primero, se determinará el grado en que la relevancia o no relevancia de un término cualquiera T_i afecta a nuestra creencia sobre la relevancia del documento. Esta información se representará mediante un conjunto de m_j probabilidades condicionadas $p(D_j \mid T_i)$, es decir, $p(d_j \mid t_i)$ y $p(d_j \mid \bar{t}_i), i = 1,\ldots,m_j$ (los valores $p(\bar{d}_j \mid t_i)$ y $p(\bar{d}_j \mid \bar{t}_i)$ se determinan por dualidad). El segundo paso consiste en determinar un mecanismo para combinar las m_j probabilidades condicionadas individuales, es decir:

$$fp(\mathbf{d}_j, <\mathbf{t_1},\ldots,\mathbf{t_{m_j}}>) = \bigotimes_{i=1}^{m_j} p(\mathbf{d}_j \mid \mathbf{t_i}),$$

siendo $\bigotimes$ el mecanismo de combinación.

En este sentido, son dos los criterios que utilizaremos como guía para determinar los valores de probabilidad: el primero, se basa en considerar que el conocer un término del documento es relevante hace que nuestra creencia sobre la relevancia del documento sea mayor que si se sabe que éste es no relevante. Establece, por tanto, que $p(d_j \mid \bar{t}_i) \leq p(d_j \mid t_i)$. El segundo que estipula la existencia de una influencia positiva cuando varios términos relevantes actúan conjuntamente. Así, si conocemos para un documento D_j la probabilidad de relevancia del documento dado que un término es relevante, $p(d_j \mid t_i)$, entonces el conocer que cualquier otro término t_l es también relevante, incrementará nuestra creencia en la relevancia del documento. Formalmente:

$$p(d_j \mid t_i,t_l) \geq max\{p(d_j \mid t_i), p(d_j \mid t_l)\}$$

Pasemos seguidamente a estudiar las distintas funciones de probabilidad que se han diseñado. Para ello, las organizaremos según el mecanismo de combinación utilizado:

1. *Funciones de probabilidad basadas en la agregación de probabilidades.*

   Esta clase de funciones, dada una configuración $\pi(D_j) =< \mathbf{t_1},\ldots,\mathbf{t_{m_j}} >$, determina la creencia sobre la relevancia del documento mediante la suma normalizada de las distintas probabilidades individuales. Esto es:

$$p(d_j \mid \pi(D_j)) = \alpha \sum_{i=1}^{m_j} p(d_j \mid \mathbf{t_i}), \quad (3.1)$$

   donde α es una constante de normalización, que se calcula como la suma de las probabilidades cuando todos los términos se consideran relevantes, es decir,

$$\alpha = \frac{1}{\sum_{i=1}^{m_j} p(d_j \mid t_i)}$$

Pasaremos seguidamente a detallar las distintas funciones de probabilidad diseñadas de acuerdo con este enfoque, aclarando antes que notaremos por $R_{\pi(D_j)}$ y $\bar{R}_{\pi(D_j)}$ el conjunto de términos que son relevantes y no relevantes, respectivamente, en la configuración $\pi(D_j)$.

- *fp3*: consideramos que todos los términos, cuando se instancian a relevantes, contribuyen en el mismo grado a la relevancia del documento, $p(d_j \mid t_i) =$ constante . Si el término se instanciara a no relevante, nuestra creencia sobre la relevancia del documento es nula. Así,

  $$p(d_j \mid t_i) = \text{ constante y } p(d_j \mid \bar{t}_i) = 0$$

  Por tanto, considerando la ecuación (3.1), la probabilidad $p(d_j \mid \pi(D_j))$ representa la proporción de términos que son considerados relevantes en la configuración $\pi(D_j)$ del total de términos que indexan el documento:

  $$p(d_j \mid \pi(D_j)) = \frac{\mid R_{\pi(D_j)} \mid}{m_j} \tag{3.2}$$

- *fp2*: se considera que la instanciación de un término a relevante afecta a nuestra creencia sobre la relevancia del documento de una manera proporcional a la calidad del término medida mediante su peso $tf \cdot idf$, es decir, $p(d_j \mid t_i) \propto tf_{ij} \cdot idf_i$.

  Al igual que en el caso anterior, la instanciación de un término a no relevante anula nuestra creencia sobre la relevancia del documento, es decir, $p(d \mid \bar{t}_i) = 0$. Con estos valores, dada una configuración $\pi(D_j)$, la probabilidad de que un documento sea relevante se obtiene mediante:

  $$p(d_j \mid \pi(D_j)) = \frac{\sum_{T_i \in R_{\pi(D_j)}} tf_{ij} \cdot idf_i}{\sum_{T_i \in D_j} tf_{ij} \cdot idf_i} \tag{3.3}$$

  Esta expresión permite calcular $p(d_j \mid \pi(D_j))$ considerando las medidas de similitud clásicas utilizadas en R.I. Esta función obtiene valores de la probabilidad equivalentes a considerar un producto escalar normalizado entre un vector documento, que almacene los valores $tf \cdot idf$ para cada término del documento, y un vector que representa la configuración $\pi(D_j)$ a evaluar, donde los términos que son relevantes se representan mediante 1 y los no relevantes a 0 [2].

  Siguiendo esta filosofía, se han considerado distintas modificaciones a la forma en que se codifica la información del vector configuración. En concreto, podemos considerar la función de probabilidad $fp8$, donde se almacena el idf para los términos que son relevantes y 0 para los que no lo son. Así, cada probabilidad condicionada individual será $p(d_j \mid t_i) \propto tf_{ij} \cdot idf_i^2$, quedando finalmente:

[2] En este caso, calcular $p(d_j \mid t_i)$ es equivalente a considerar la configuración $(0,0,0,0,\ldots,1,0,\cdots 0)$, es decir, un 1 en la i-ésima posición.

$$p(d_j \mid \pi(D_j)) = \frac{\sum_{T_i \in R_{\pi(D_j)}} tf_{ij} \cdot idf_i^2}{\sum_{T_i \in D_j} tf_{ij} \cdot idf_i^2} \quad (3.4)$$

La otra alternativa aparece cuando el vector configuración también se almacena el $tf \cdot idf$, dando lugar a la función de probabilidad $fp4$, donde $p(d_j \mid t_i) \propto (tf_{ij} \cdot idf_i)^2$ y

$$p(d_j \mid \pi(D_j)) = \frac{\sum_{T_i \in R_{\pi(D_j)}} (tf_{ij} \cdot idf_i)^2}{\sum_{T_i \in D_j} (tf_{ij} \cdot idf_i)^2} \quad (3.5)$$

2. *Métodos basados en una puerta OR ruidosa.*

   Antes de pasar a describirlos, vamos a introducir brevemente el concepto de *interacción disyuntiva*, más conocido como *puerta OR ruidosa* [Pea88, Jen96]. Cuando una variable x tiene varios padres $\Pi(x)$, se debe estimar $p(\mathbf{x} \mid \pi(x))$ para cada configuración de valores $\pi(x)$ de los padres de dicha variable. Puede darse el caso de que existan distribuciones de probabilidad que sean difíciles de obtener, bien porque se dan pocos casos en la base de datos, bien porque el experto no es capaz de hacer una estimación de las mismas. En ese caso, se buscaría una forma de, a partir de unas distribuciones de probabilidad cuya estimación fuera viable, de poder calcular las que realmente hacen falta.

   Desde el punto de vista de la causalidad, una interacción disyuntiva entre las diferentes causas de un efecto (padres de una variable) se da cuando cualquiera de ellas puede originar el efecto de manera independiente y sin que se produzca ninguna pérdida de creencia sobre ella cuando ocurren a la vez varias causas. Se deben dar dos condiciones para su aplicación:

   - Un suceso E es falso (la probabilidad de E es cero) cuando todas las causas de E también lo son.
   - Si un suceso E es consecuencia de una de las dos condiciones C_1 o C_2, entonces el mecanismo que inhibe la ocurrencia de E a pesar de la aparición de la condición C_1 es independiente del mecanismo que inhibe E cuando se da C_2.

   Sea x una variable binaria que representa un efecto que tiene como padres en una red bayesiana al conjunto de variables $\Pi(x) = \{u_1, u_2, \ldots, u_n\}$, también binarias, que actúan como causas de x. Además, existen una serie de inhibidores, $I_1, I_2, \ldots I_n$ que interfieren en la relación normal de $\Pi(x)$ con x (esta información no se representa explícitamente en la red bayesiana, sino que se expresa implícitamente en el conjunto de distribuciones de probabilidad $P(x \mid \Pi(x))$.

   Este modelo de puerta OR ruidosa se puede representar gráficamente como se muestra en la figura 3.3, donde cada causa individual es una entrada de una puerta *AND*. La otra entrada será la negación de su correspondiente inhibidor. Cada una de las salidas de la

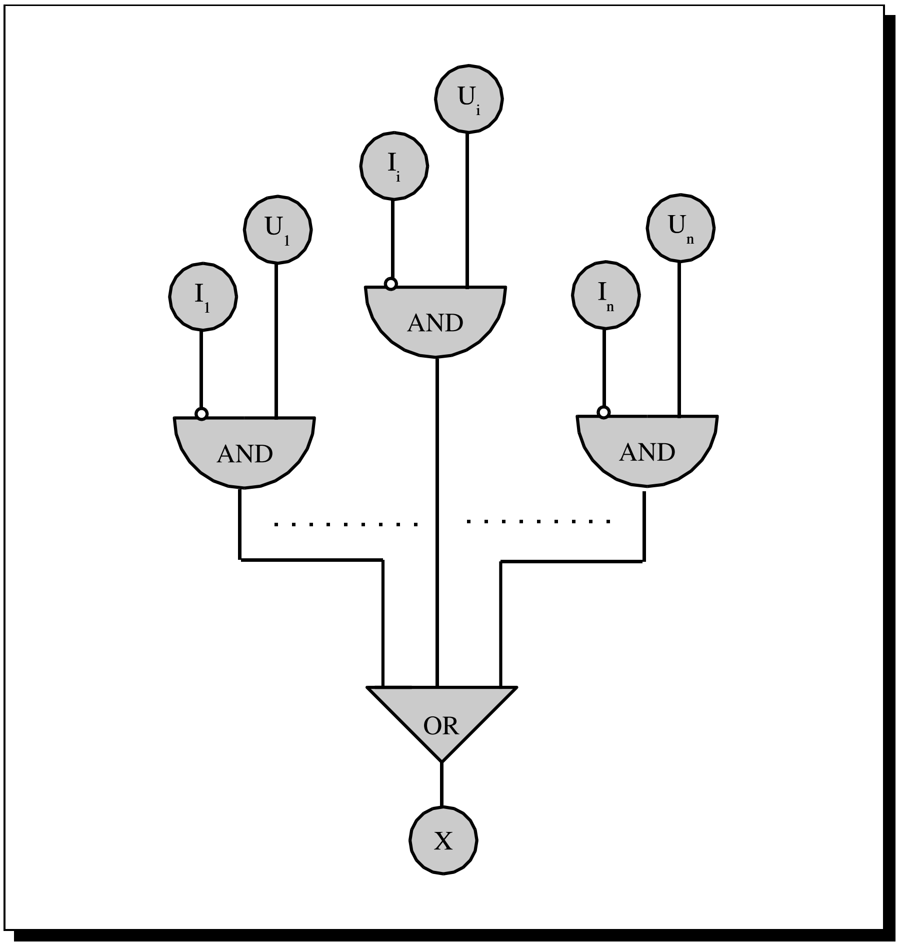


Figura 3.3.: Representación de una puerta OR ruidosa.

puertas *AND* se configurará como entrada para una puerta *OR*. La salida de esta última será el efecto x.

Si u_i es la única causa de x que está activa, entonces x será verdadero si, y sólo si, el inhibidor asociado con u_i permanece inactivo. Por tanto, si la probabilidad de que el inhibidor esté activo es q_k, esto es, $p(i_k) = q_k$, tendremos que $p(x \mid u_i, \bar{u}_k, \forall k \neq i) = 1 - q_i$. Así, $c_i = 1 - q_i$ será la probabilidad de que una causa u_i pueda hacer que el efecto x se produzca.

Sea $T_{\pi(x)} = \{i \mid \mathbf{u_i} = u_i\}$, donde $\pi(x)$ es una configuración de valores que pueden tomar las variables pertenecientes a $\Pi(x)$. La distribución de probabilidad condicionada de x respecto a una configuración $\pi(x)$ de sus padres se puede calcular aplicando la siguiente expresión:

$$p(\mathbf{x} \mid \pi(x)) = \begin{cases} \prod_{i \in T_{\pi(x)}} q_i & si \quad \mathbf{x} = \bar{x} \\ 1 - \prod_{i \in T_{\pi(x)}} q_i & si \quad \mathbf{x} = x \end{cases} \tag{3.6}$$

Trasladando esta idea al ámbito de la R.I. donde nos movemos, la variable efecto x es un documento D_j y sus causas, el conjunto de variables término padres, es decir, $\Pi(D_j)$. En este caso, la probabilidad de que el inhibidor esté activo coincide con la probabilidad con que el documento no es relevante dado que el término sí lo es, es decir, $p(\bar{d} \mid t_i) = 1 - p(d \mid t_i)$.

Las dos funciones que se han diseñado siguiendo este enfoque son las siguientes:

- *fp5*: este método calcula $p(d_j \mid t_i)$ sobre la base del $tf \cdot idf$, tratando de recoger la calidad del término a lo largo de toda la colección y dentro del propio documento. Esta medida se normaliza por el máximo $tf \cdot idf$ de todos los términos del documento. Al valor normalizado lo notaremos como $Norm(tf \cdot idf)$, tomando sus valores en el intervalo $[0,1]$. Al estudiar esos valores en los documentos de las colecciones analizadas, observamos que eran bastante próximos a 1. Esto chocaba con nuestra intuición de que $p(d_j \mid t_i)$ no debería ser muy alta. Por tanto, decidimos aplicar factores correctores: el primero consiste en un cambio de escala al dividirlo por $\log(N+1)$, y el segundo trata de recoger la idea de considerar que para un documento que haya sido indexado por muchos términos, conocer que sólo un término es relevante da menos información que el mismo hecho cuando el documento ha sido indexado por pocas palabras clave. Por tanto, la expresión para un único término es la siguiente:

  $$p(d_j \mid t_i) = \frac{Norm(tf_{ij} \cdot idf_i)}{log(N+1)log(m_j+1)} \text{ y } p(d_j \mid \bar{t}_i) = 0$$

  Finalmente, para obtener $p(D_j \mid \pi(D_j))$, las probabilidades individuales se combinan utilizando la expresión (3.6):

  $$p(d_j \mid \pi(D_j)) = 1 - \prod_{T_i \in R_{\pi(D_j)}} (1 - p(d_j \mid t_i)) \tag{3.7}$$

- *fp1*: son dos las diferencias que existen con el esquema anterior: por un lado, considera un idf normalizado por el máximo en lugar del $tf \cdot idf$ normalizado del anterior, tratando de recoger la idea de que cuanto más aparece un término en la colección, menos probable será que los documentos que lo contienen sean relevantes. Por otro lado, que $p(d_j \mid \bar{t}_i)$ no es cero, sino que se le asigna un valor constante para todos los documentos, pero muy próximo a cero. Además, se incluyen los mismos factores correctores que en la anterior función. La forma de calcular $p(d_j \mid t_i)$ queda:

  $$p(d_j \mid t_i) = \frac{Norm(idf_i)}{log(N+1)log(m_j+1)}$$

  Finalmente nos queda exponer la función de probabilidad también basada en una puerta OR ruidosa, adaptada para considerar que $p(d_j \mid \bar{t}_i)$ es distinta de cero. Así,

$p(d_j \mid t_i) = p_i$ y $p(d_j \mid \bar{t}_i) = q_i$; entonces, si consideramos que todos los términos son relevantes:

$$p(d_j \mid t_1, \ldots, t_{m_j}) = 1 - \prod_{i=1}^{m_j}(1 - p_i)$$

Cuando todos son no relevantes:

$$p(d_j \mid \bar{t}_1, \ldots, \bar{t}_{m_j}) = \prod_{i=1}^{m_j} q_i$$

En este caso, se puede observar que en realidad no corresponde esta expresión a una puerta *OR*, sino a una especie de “1 - puerta *OR*". Y finalmente, cuando algunos de los términos son relevantes $(t_1, \ldots, t_{m_j})$ y el resto $(\bar{t}_1, \ldots, \bar{t}_{m_j})$ no lo son, se aplica la siguiente expresión, donde se combinan los dos casos anteriormente citados:

$$p(d_j \mid t_1, \ldots, t_h, \bar{t}_{h+1}, \ldots, \bar{t}_{m_j}) = 1 - \prod_{i=1}^{h}(1 - p_i) + \prod_{k=h+1}^{m_j} q_k \prod_{i=1}^{h}(1 - p_i)$$

3. *Métodos basados en la medida de similitud del coseno.*

   El fundamento de estos métodos radica en medir la similitud entre el documento D_j y la configuración $\pi(D_j)$, utilizando como base una medida ampliamente utilizada en R.I.: la medida del coseno, que obtiene el coseno del ángulo que forman el documento y la configuración que se desea evaluar. Esta medida devuelve valores de probabilidad en el intervalo $[0, 1]$. Según esta idea, hemos diseñado las siguientes funciones de probabilidad:

   - *fp10*: en esta función de probabilidad, el documento D_j se representa por el vector $(tf_{1j} \cdot idf_1, \ldots, tf_{m_j j} \cdot idf_{m_j})$, mientras que la configuración contendría los idf de los términos que fueran relevantes y 0 para los que no lo fueran. Así, la estimación de la probabilidad de que el documento sea relevante dada una configuración se puede ver como una expresión al estilo de la que calcula el coseno del ángulo que forman ambos vectores:

   $$p(d_j \mid \pi(D_j)) \propto \frac{\sum_{T_i \in R_{\pi(D_j)}} tf_{ij} \cdot idf_i^2}{\sqrt{\sum_{T_i \in D_j} tf_{ij} \cdot idf_i^2}}$$

   - *fp6*: en esta función, la configuración queda formada por el $tf \cdot idf$ de los términos que aparecen en ella como relevantes y 0 para los que figuran como no relevantes. Así, la probabilidad condicionada del documento toma la forma siguiente:

   $$p(d_j \mid \pi(D_j)) \propto \frac{\sum_{T_i \in R_{\pi(D_j)}} (tf_{ij} \cdot idf_i)^2}{\sqrt{\sum_{T_i \in D_j} (tf_{ij} \cdot idf_i)^2}}$$

Por tanto, se han diseñado dos variedades de funciones de probabilidad basadas ambas en la medida del coseno, con la peculiaridad de que también se pueden incluir en los métodos de estimación basados en agregación de probabilidades; ya que *fp10* se puede ver cómo $p(d_j \mid t_i) \propto tf_{ij} \cdot idf_i^2$ y, análogamente, *fp6* se puede expresar como $p(d_j \mid t_i) \propto (tf_{ij} \cdot idf_i)^2$.

3.3. La red aumentada.

En la red simple consideramos los términos de la colección independientes entre sí. Este hecho puede verse como una suposición bastante estricta a la hora de modelar el problema, pues es fácil pensar en la existencia de relaciones de dependencia entre términos. Por ejemplo, si conocemos que el término "bayesiano" es relevante a una consulta, entonces podemos estar interesados en aquellos documentos que estén indexados por la palabra "probabilidad".

Por tanto, y partiendo de una red simple, el paso natural para obtener un modelo más preciso es olvidarse del criterio inicial de independencia entre términos, permitiendo al modelo incorporar las relaciones de dependencia más importantes entre los términos de la colección. La red simple dará paso a un nuevo modelo de red bayesiana para la recuperación de información, al que hemos dado en llamar la *red bayesiana documental basada en la subred de términos aumentada*, o de forma abreviada, *la red aumentada*. Con este cambio se espera obtener mejores resultados en la recuperación, pues se está utilizando un modelo más preciso que la red simple y, por tanto, más cercano a la realidad.

En este nuevo modelo vamos a seguir estableciendo la diferencia entre la subred de términos y la subred de documentos. La unión de ambas redes se hará, al igual que se hizo en la red simple, por medio de arcos que saldrán de los nodos término y apuntarán a aquellos nodos documento que los contienen.

Centrándonos inicialmente en la subred de documentos, optaremos por una estructura idéntica a la que se ha tomado en la red simple, es decir, una red donde los nodos documento están completamente aislados, reflejando la suposición de independencia entre nodos documento dado que conocemos los nodos término que los indexan.

En cuanto a la subred de términos, la topología subyacente debe permitir la inclusión de las relaciones entre términos. Con la idea de automatizar el proceso, nos planteamos el uso de un algoritmo de aprendizaje. Teniendo en cuenta la problemática que representa el aprendizaje automático de las relaciones entre términos de una colección, ya detallados en el capítulo anterior, restringiremos la estructura del modelo a un poliárbol. En la figura 3.4 podemos ver gráficamente la nueva topología de la red bayesiana, donde con líneas discontinuas se representan los arcos que conectan términos.

El algoritmo de aprendizaje utilizado para generar el poliárbol es el mismo que el que se explicó en el capítulo anterior para construir el tesauro.

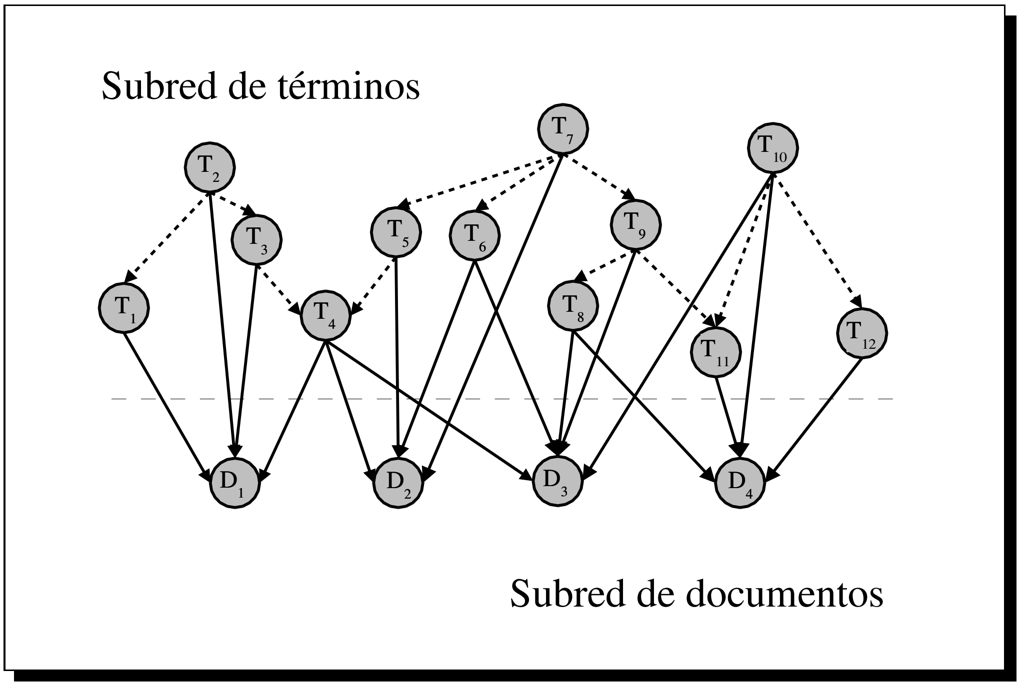


Figura 3.4.: Red bayesiana documental formada por la subred de términos aumentada.

3.3.1. Estimación de la información cuantitativa.

Pasaremos seguidamente a presentar cómo se calculan las distintas distribuciones de probabilidad alojadas en los nodos de la red.

Con respecto a los nodos documento, vamos a usar la misma técnica que utilizamos en la red simple para evitar tener que estimar y, posteriormente, almacenar las distribuciones de probabilidad condicionada en los nodos documento: usar funciones de probabilidad, que serán las mismas que las presentadas para el modelo simple.

En este nuevo modelo, los nodos término pueden interpretar dos papeles: por un lado, si no tienen padres, sólo almacenarán una distribución de probabilidad marginal. En este nuevo modelo el número de nodos raíces se ve reducido considerablemente. De igual forma que con el modelo anterior, se pretende estudiar si los resultados de la recuperación son sensibles al tipo de cálculo elegido, por lo que se considerarán los estimadores $pp1$, $pp2$, $pp3$ y $pp4$. Por otro lado, al estar trabajando con un poliárbol, habrá otro conjunto de nodos término que sí tienen padres. En este caso, al igual que ocurre con los nodos documento, será necesario estimar las distribuciones de probabilidad condicionales. Nos vamos a centrar en la estimación de estas distribuciones de probabilidad, es decir, $P(T_i \mid T_1, \ldots, T_p)$, donde $T_1, \ldots, T_p$ son los padres del nodo T_i

Dado un nodo T_i y el conjunto de sus padres, $\Pi(T_i)$, tres son los métodos que hemos empleado para realizar esta estimación:

- *pc-mv (estimador de máxima verosimilitud)*. Aunque ya fue introducido en el capítulo 2 como estimador base para las distribuciones condicionadas de los nodos término del poliárbol, pasamos a recordarlo y a poner un ejemplo para ver su funcionamiento y dónde se manifiestan algunos posibles problemas.

  El estimador en cuestión se corresponde con la siguiente expresión:

$$p(t_i \mid \pi(T_i)) = \frac{n(< t_i, \pi(T_i) >)}{n(\pi(T_i))} \text{ y } p(\bar{t}_i \mid \pi(T_i)) = 1 - p(t_i \mid \pi(T_i))$$

  Ejemplo 1*:* Supongamos que tenemos tres variables aleatorias que representan a los términos *LENGUAJE*, *PROGRAMA* y *COMPILADOR* de una colección cualquiera sobre informática, y referenciados, respectivamente, como L, P y C. Supongamos, también, que la relación entre ellos se corresponde gráficamente con la figura 3.5.

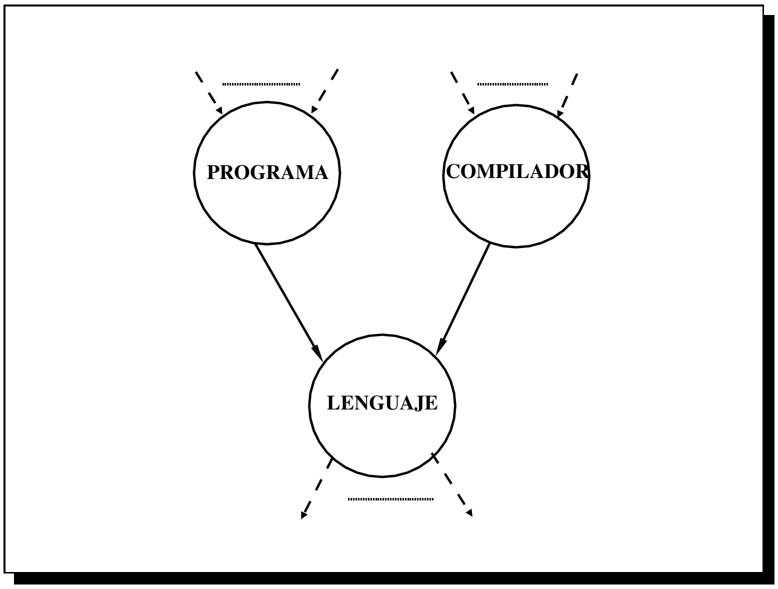


Figura 3.5.: Un nodo término con dos padres en la subred aumentada.

  Para calcular la distribución de probabilidad condicionada del nodo L dados sus padres, partimos de la matriz de frecuencias obtenida directamente de la base de datos documental contando las veces en que aparecen todas las combinaciones posibles de los tres términos:

$$n(L,P,C) = \begin{array}{c|cccc} & \overline{p},\overline{c} & \overline{p},c & p,\overline{c} & p,c \\ \hline \overline{l} & 33 & 0 & 2 & 10 \\ l & 5 & 0 & 0 & 30 \end{array}$$

  A esta matriz le aplicamos la expresión *pc-mv* y obtenemos la distribución de probabilidad que estamos persiguiendo. Por ejemplo:

$$p(L \mid \bar{p},\bar{c}) = \begin{array}{c|ccc} & & \bar{p},\bar{c} & \\ \hline \bar{l} & \frac{33}{38} & = & 0{,}87 \\ l & \frac{5}{38} & = & 0{,}13 \end{array} \text{ y } p(L \mid p,c) = \begin{array}{c|ccc} & & p,c & \\ \hline \bar{l} & \frac{10}{40} & = & 0{,}25 \\ l & \frac{30}{40} & = & 0{,}75 \end{array}$$

El uso del estimador de máxima verosimilitud para realizar la estimación de las matrices de probabilidad condicionada presenta dos problemas fundamentales [Cam98b]:

1. La dispersión de datos: el estimador puede no estar definido cuando el número de datos es cero. Esta situación es frecuente en una colección grande, con un número elevado de términos entre los que existe una gran cantidad de patrones que no aparecen en la base de datos. Así, en el ejemplo anterior, $p(l \mid \bar{p},c) = 0/0$.
2. El sobreajuste: si el número de datos disponibles es pequeño, el estimador puede ajustarse demasiado a los datos disponibles sin que represente realmente el verdadero valor de la probabilidad. Al calcular la distribución de probabilidad condicionada, habrá muchas probabilidades iguales a cero o muy próximas (análogamente las habrá iguales a uno o muy cercanas), originando una matriz de muy poca calidad. En el ejemplo, este es el caso de $p(\bar{l} \mid p,\bar{c}) = 1$.

Para solucionar estos problemas, nos hemos planteado el uso de:

- *pc-eb* (estimadores bayesianos) [Goo65]:

  1. *eb1*: este estimador está basado en la *ley de la sucesión de Laplace* [Goo65], la cual establece que si en una muestra de N casos encontramos k casos que verifican una determinada propiedad Q (por ejemplo, que el valor de una variable sea uno dado), entonces la probabilidad de que el siguiente caso que observemos exhiba la misma propiedad es $(k+1)/(N+ \mid Q \mid)$, donde $\mid Q \mid$ representa las alternativas posibles que se consideran para la propiedad Q (por ejemplo, el número de valores distintos que puede tomar una variable).

     Según esta ley anterior, la forma de estimar la probabilidad condicionada de que un término sea relevante dada una configuración de sus padres es la siguiente:

     $$p(t_i \mid \pi(T_i)) = \frac{n(< t_i, \pi(T_i) >) + 1}{n(\pi(T_i)) + \mid T_i \mid}$$

     siendo $\mid T_i \mid$ el número de valores que puede tomar T_i. Si la muestra es muy grande, el valor devuelto por este estimador tenderá a acercarse al calculado por *pc-mv*, mientras que cuando sea muy pequeña, se parecerá a una distribución uniforme. La razón por la cual se puede calificar de estimador bayesiano es porque se parte de cierta información a priori y se actualiza a la luz de nuevos resultados. En este caso, la información a priori es uniforme.

     Ejemplo 1 *(continuación)*. Aplicando *eb*1 obtendríamos la siguiente matriz:

$$p(L \mid P,C) = \begin{array}{c|cccc} & \overline{p},\overline{c} & \overline{p},c & p,\overline{c} & p,c \\ \hline \overline{l} & \frac{33+1}{38+2} = 0{,}85 & \frac{0+1}{0+2} = 0{,}5 & \frac{2+1}{2+2} = 0{,}75 & \frac{10+1}{40+2} = 0{,}27 \\ l & \frac{5+1}{38+2} = 0{,}15 & \frac{0+1}{0+2} = 0{,}5 & \frac{0+1}{2+2} = 0{,}25 & \frac{30+1}{40+2} = 0{,}73 \end{array}$$

Como se puede apreciar, las probabilidades que antes eran cero ahora dejan de serlo tomando valores algo mayores.

2. *eb2*: en la misma línea que el anterior, aunque más general, el también denominado *m-estimación* [Ces90, CB91] tiene la forma:

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

siendo s un parámetro que se puede interpretar como el tamaño muestral necesario para estimar la distribución a priori.

Este segundo método presenta dos variantes a la hora de llevarlo a cabo: una primera en la que se aplica la expresión $eb2$ para estimar todas las distribuciones de probabilidad condicionada, con objeto de suavizarlas de forma global; y una segunda, en la que sólo se aplica para estimar la distribución $p(\mathbf{t_i} \mid t_1, \cdots, t_p)$, correspondiente a la configuración donde todas las variables padre toman el valor de relevante. La razón para esta conducta es que esta probabilidad es la más importante, ya que nos da el valor de relevancia de un término cuando todos su padres son relevantes, valor que normalmente suele ser muy cercano a cero y que de esta forma se suaviza. El resto de distribuciones de probabilidad se estiman utilizando el estimador *pc-mv*.

Un estudio experimental detallado sobre el comportamiento de estos métodos de estimación bayesiana aplicados a la colección ADI puede encontrarse en [CFH00].

- *pc-J (estimador basado en el coeficiente de Jaccard)*:

Esta fórmula se basa en la medida de similitud de Jaccard [Rij79], 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 forman la unión de ambos conjuntos, es decir,

$$\frac{\mid X \cap Y \mid}{\mid X \cup Y \mid}$$

Esta medida, que ya fue utilizada por Savoy para calcular las probabilidades condicionadas en su modelo [SD91], queda adaptada en la siguiente expresión al nuestro:

$$p(\bar{t}_i \mid \pi(T_i)) = \frac{n(< \bar{t}_i, \pi(T_i) >)}{n(< \bar{t}_i >) + n(\pi(T_i)) - n(< \bar{t}_i, \pi(T_i) >)} \text{ y } p(t_i \mid \pi(T_i)) = 1 - p(\bar{t}_i \mid \pi(T_i))$$

Se puede observar que en primer lugar se estima $p(\bar{t}_i \mid \pi(T_i))$ y $p(t_i \mid \pi(T_i))$ se determina por dualidad. Básicamente, la razón por la que procedemos así se debe a que si se utiliza

Jaccard para calcular los dos valores de probabilidad, no se obtendría una distribución de probabilidad, es decir, si sumamos individualmente los valores, no obtenemos la unidad. Además, el hecho de aplicar este estimador al caso en que la variable condicionada es no relevante y luego obtener la del caso relevante, y no al revés, se debe a que este estimador genera valores de probabilidad más bajos que el de máxima verosimilitud, reduciendo así la probabilidad de no relevancia y aumentando la de relevancia.

Ejemplo 1 *(continuación)*. Estimando con *pc-J* a partir de la matriz de frecuencias se obtiene:

$$p(L \mid P,C) = \begin{array}{c|cccc} & \overline{p},\overline{c} & \overline{p},c & p,\overline{c} & p,c \\ \hline \overline{l} & \frac{33}{50}=0{,}66 & \frac{0}{45}=0 & \frac{2}{45}=0{,}04 & \frac{10}{75}=0{,}13 \\ & 0{,}34 & 1{,}0 & 0{,}96 & 0{,}87 \end{array}$$

3.4. La red mixta.

El hecho de permitir la existencia de enlaces entre términos se ha mostrado deseable para nuestro S.R.I. La idea subyacente al uso de estas relaciones de términos es permitir que también se recuperen aquellos documentos que, sin estar indexados por los términos de la consulta Q, lo están por otros términos estrechamente relacionados con los términos de la consulta. Sin embargo, un estudio más detallado de los enlaces existentes en la red nos muestra algunas peculiaridades que pueden afectar a la capacidad recuperadora del sistema.

Por ejemplo, supongamos dos términos, T_1 y T_2, que únicamente aparecen en el documento D_3. En este caso, la dependencia existente entre estos dos términos es fuerte (siempre que aparece uno, aparece el otro), por lo que, probablemente, el algoritmo de aprendizaje añada a la red que está construyendo una arista uniéndolos. Sin embargo, cuando utilizamos el modelo en procesos de recuperación, la existencia de este tipo de enlaces será poco significativa, pues pueden no ayudar a favorecer el traspaso de información entre términos que pertenezcan a documentos distintos en el momento de la propagación.

Para solucionar este problema, nos planteamos realizar una selección de los términos previa al proceso de aprendizaje. Este proceso proporcionará además efectos colaterales beneficiosos relacionados con la eficiencia del proceso de aprendizaje de la estructura, ya que se reduce la dimensión del problema, disminuyendo así el tiempo de la fase de aprendizaje. Esta ventaja es tan importante que, incluso obteniendo una pequeña pérdida en la capacidad recuperadora, este nuevo modelo sería totalmente válido. Esta idea de disminuir el tamaño del problema por resolver ya ha sido puesta en práctica en otro ámbitos de la R.I. basada en redes bayesianas, pues Sahami [Sah98] desarrolló un método de selección de características para reducir el número de términos con los que llevar a cabo tareas posteriores de clasificación.

La idea que yace en este nuevo modelo es la búsqueda de aquellos términos que se consideren mejores o superiores bajo algún criterio. Con estos términos se procedería a aprender un poliárbol, el cual deberá ser de mayor calidad que si se utilizan todos los términos, como ocurre en la red aumentada.

El resto de términos que se quedan fuera del poliárbol no se eliminarían de la subred de términos: se añadirán a dicha subred de forma totalmente aislada, ya que no habrá ningún enlace ni entre ellos ni con los términos incluidos en el poliárbol. Esta estructura de la subred de términos se puede ver como una estructura mixta, ya que almacena en el mismo grafo un poliárbol con los mejores términos y una estructura simple con el resto. En la figura 3.6 se muestra gráficamente el proceso expuesto anteriormente.

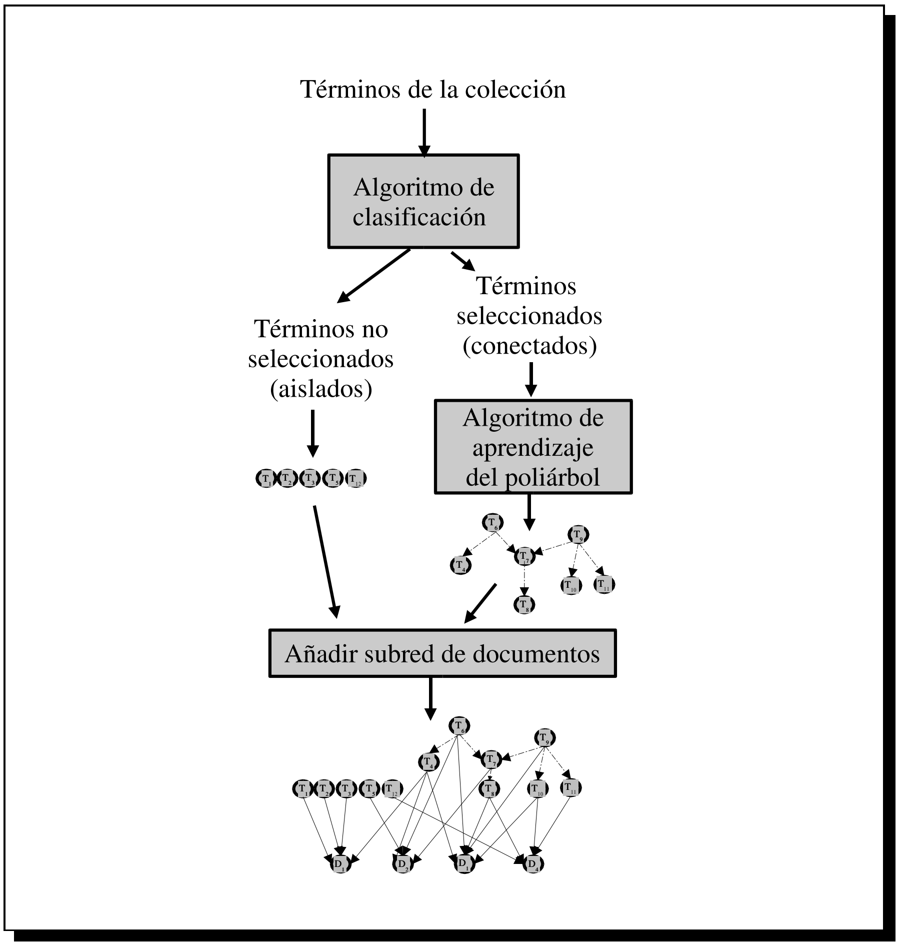


Figura 3.6.: Proceso de construcción de la red mixta.

La subred de documentos permanece intacta y la conexión entre nodos término y documentos se lleva a cabo de igual manera que se hacía en las redes simple y aumentada. A este modelo le daremos el nombre de *red bayesiana documental basada en la subred de términos mixta*, o lo que es lo mismo en su forma abreviada, *la red mixta*, cuya topología se presenta en la figura 3.7.

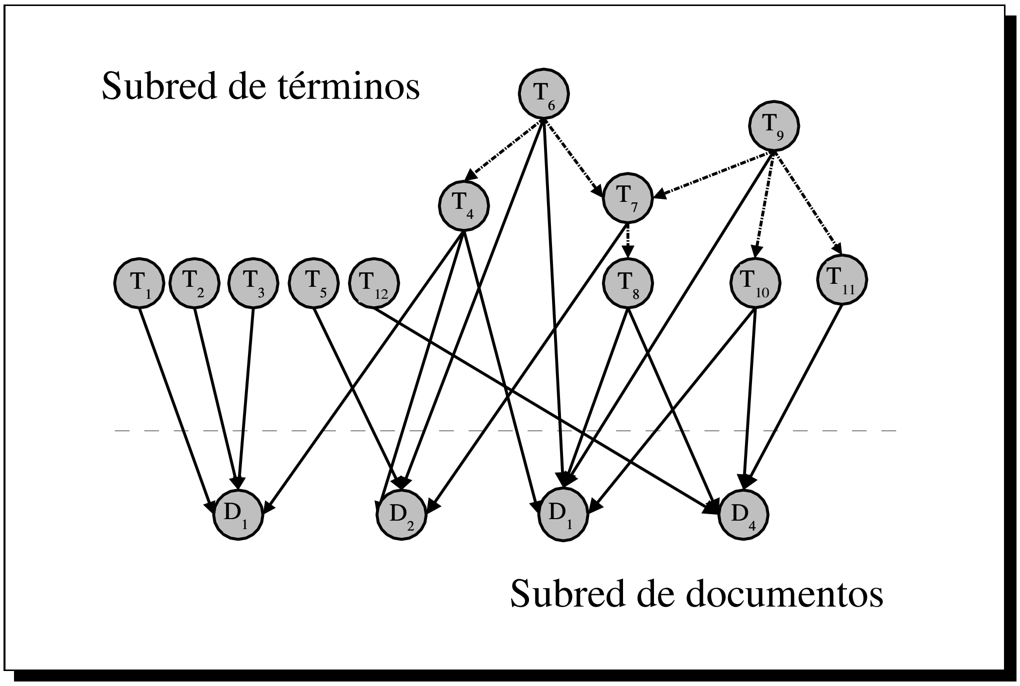


Figura 3.7.: Red bayesiana documental formada por la subred de términos mixta.

3.4.1. Selección de términos.

El objetivo que nos proponemos alcanzar es clasificar los términos en dos grupos: los que pertenecen al poliárbol, que notaremos como $\mathcal{C}$ (Conectados), y los que no, que designaremos como $\mathcal{A}$ (Aislados).

Una primera alternativa podría ser aquélla basada en los trabajos de Luhn sobre análisis automático de textos [Rij79], que ya comentamos en el capítulo 1. Este investigador desarrolló una técnica basada en la ley de Zipf para encontrar aquellas palabras más útiles para representar el contenido de un documento, es decir, seleccionar los términos de la indexación. El problema fundamental que se nos planteaba si seguíamos la idea de Luhn era determinar los valores de los cortes superior e inferior, ya que el objetivo principal es realizar de manera automática el proceso de selección.

La propuesta que nosotros hacemos se basa en combinar la información que aportan el valor de discriminación del término y la frecuencia documental inversa. Pasaremos seguidamente a describir cada una de estas dos medidas individualmente para luego comentar cómo se han combinado y la forma en que se ha realizado la selección de términos.

- *Valor de discriminación del término (vdt).* [SM83]. Mide el grado con el que un término es capaz de distinguir unos documentos de otros. Se basa en el cálculo de una medida de similitud entre documentos, S, en base a los términos comunes que poseen , de tal forma que cuando dos documentos D_j y D_k sean iguales, $S(D_j, D_k) = 1$ y cuando ambos no tengan términos en común, entonces $S(D_j, D_k) = 0$. El valor de discriminación del término i-ésimo se obtendrá como la diferencia entre la similitud media entre los documentos, $\bar{S}$, y la similitud media de dichos documentos tras eliminar el término i-ésimo, $\bar{S}_i$, es decir, $vdt = \bar{S}_i - \bar{S}$. Así, para un término que aparezca en muchos documentos de la colección, el hecho de quitarlo originará que la similitud media entre los documentos, $\bar{S}$, se decremente, lo que implica que los documentos serán más diferentes entre sí. Por el contrario, si se elimina un término que aparece pocas veces, el efecto es el contrario, es decir, los documentos se acercan más y su similitud media se hace mayor.

Algoritmo 3.1 Cálculo del valor de discriminación de los términos de la colección.

entradas: colección de documentos y glosario de la misma.
salidas: valor de discriminación de cada término del glosario.
Calcular la similitud de cualesquiera dos documentos D_j y D_k, $j \neq k$.
Obtener la similitud media, $\bar{S}$, la cual refleja la densidad del espacio de documentos.
para cada término, T_i, de la colección **hacer**
    Calcular la similitud de cualesquiera dos documentos D_j y D_k, $j \neq k$, pero quitando de éstos el término T_i en caso de que aparezca en ellos.
    Obtener la similitud media, $\bar{S}_i$, a partir de las nuevas similitudes, representando la densidad del espacio de documentos al quitar el término T_i.
    El valor de discriminación del término T_i será:

$$vdt(T_i) = \bar{S}_i - \bar{S}$$

fin para

Si ordenamos todos los términos de forma decreciente según su valor vdt, calculado siguiendo los pasos del algoritmo 3.1, podremos clasificarlos en tres grupos:

- Términos con $vdt > 0$: son los buenos discriminadores ya que decrementan la densidad del espacio documental.
- Términos con $vdt = 0$: indiferentes con respecto al poder discriminador. Su eliminación deja el espacio casi inalterado.
- Términos con $vdt < 0$: hacen que los documentos sean más similares, y por tanto, son malos discriminadores.

Los términos con un valor discriminador menor que cero se corresponden con aquellos que tienen una frecuencia de aparición muy elevada. Por el contrario, los buenos discriminadores son los que poseen un valor mayor que cero, lo que se traduce en una frecuencia media, hecho que apoya la teoría de Luhn.

Centrándonos en detalles de la implementación de este algoritmo, cabe destacar que la medida de similitud usada ha sido la del coseno [SM83]. Además, si se implementa directamente el algoritmo 3.1 que calcula el valor vdt para cada término, el proceso se ejecutará en tiempo perteneciente a $O(NM^2)$. Una variación que tiene por objeto reducir su complejidad algorítmica pasa por calcular un documento centroide: un documento que contiene todos los términos de la colección, siendo el peso de cada término la media de los pesos de los documentos en los que aparece. En vez de calcular las similitudes de todos los documentos entre sí, se calculará la distancia de todos los documentos al centroide, obteniéndose posteriormente la distancia media al centroide. Este proceso se repite quitando sucesivamente cada término de la colección. La medida *vdt* será, al igual que ocurría anteriormente, la diferencia de ambas medias, dejando el proceso con un tiempo $O(NM)$.

Por tanto, como hemos comentado, un criterio de selección de términos para su inclusión en el poliárbol consiste en considerar aquéllos con $vdt > 0$. Sin embargo, cuando aplicamos este proceso a los términos de las colecciones con las que hemos experimentado, no se encuentra un valor claro con el que poder seleccionar a los buenos discriminadores. El máximo *vdt* calculado siempre es muy próximo a cero. Además, la diferencia entre el *vdt* de cada dos términos consecutivos se podía establecer en las milésimas, hecho que dificultaba aún más la tarea de establecer un corte.

- *Frecuencia inversa del término.* La segunda alternativa consiste en utilizar la frecuencia inversa de cada término, idf, que ofrece un valor que es inversamente proporcional al número de ocurrencias del término en la colección. Así, cuanto más frecuente sea el término, menor será el valor calculado y, análogamente, cuantas menos veces aparezca en la colección, más grande será el valor obtenido. Por tanto, una vez ordenados los términos según su idf, los términos que nos interesan se sitúan en la zona central de la ordenación, descartando a los más frecuentes, por un lado, y a los menos comunes, por otro. De nuevo, nos encontramos con el problema de la determinación de los valores de corte.

Fijémonos seguidamente en el significado de las dos últimas medidas: el valor discriminador de un término ofrece una idea de la capacidad de un término para hacer que los documentos donde aparezca sean más o menos similares; la frecuencia inversa, a su vez, indica la generalidad o especificidad de un término en la colección. Lo ideal para nuestros fines sería seleccionar los términos con un valor de discriminación mayor y con una frecuencia inversa media-alta, con lo cual nos estamos asegurando que captamos los mejores términos de la colección según dos medidas diferentes. Por tanto, el método de selección de términos para el aprendizaje propuesto está basado en la medida combinada, para cada término, de su *vdt* y su idf. Con la idea de automatizar el proceso, dejamos la responsabilidad de la selección propiamente dicha a la aplicación de un algoritmo de clasificación no supervisado que sea capaz de situar cada término en una de las dos clases posibles: la clase de términos *conectados* ($\mathcal{C}$), que serán incluidos en el poliárbol, y la de términos *aislados* ($\mathcal{A}$), que serán introducidos en la subred de términos, pero dejándolos aislados. El algoritmo de clasificación utilizado para la selección de términos es el de las *k Medias* utilizando como distancia la euclídea [JMF99].

Clasificación de los términos en base a las medidas *vdt* e idf. Un algoritmo de clasificación no supervisado tiene como objetivo agrupar un conjunto de objetos, caracterizados por

Colección	Total términos	Aislados	%	Conectados	%
ADI	828	603	72.83	225	27.17
CACM	7562	6767	89.49	795	10.51
CISI	4985	4281	85.88	704	14.22
CRANFIELD	3857	3350	86.85	507	13.15
MEDLARS	7170	6198	86.44	972	13.56

Cuadro 3.1.: Distribución de los términos de las cinco colecciones en las clases $\mathcal{C}$ y $\mathcal{A}$ al aplicar el algoritmo de las *k medias*.

una serie de atributos, de acuerdo a las posibles similitudes entre ellos, generando así varias clases o grupos. Cada una de estas clases estará representada por un *centroide*, es decir, un objeto ficticio que resume el contenido de dicha clase. Los objetos que compongan un grupo serán más parecidos entre ellos que con respecto a los objetos de otros grupos. Esta similitud se mide como una función de distancia entre objetos, generalmente la distancia euclídea.

El algoritmo de las *k medias* parte inicialmente de una distribución arbitraria de objetos en las clases que se desean generar. Seguidamente calcula los centroides de dichas clases haciendo la media de cada uno de los atributos de los objetos incluidos en cada una de ellas. El próximo paso será determinar la distancia de cada objeto a cada uno de los diferentes centroides. Si alguna de estas distancias calculadas es menor que la que existe con el del centroide de la clase a la que pertenece, entonces se cambia de grupo. Este proceso se reitera hasta que los centroides de cada clase permanezcan inalterados.

En nuestro caso, los objetos serán los términos de la colección, los cuales vendrán caracterizados por dos atributos: su idf y vdt. La clasificación se pretende hacer en dos clases de términos: los que compondrán el poliárbol y los que permanecerán aislados entre ellos ($\mathcal{C}$ y $\mathcal{A}$ respectivamente). El algoritmo de las k medias que se ha utilizado es el implementado en la aplicación de tratamiento estadístico *STATGRAPHICS*.

Como consecuencia de la aplicación del algoritmo de las *k medias* a las cinco colecciones con las que estamos experimentando, los tamaños de las dos clases para cada una de ellas se muestran en la tabla 3.1. En ella vemos el total de términos de cada una de las colecciones y el número de éstos que han sido clasificados en cada clase, así como el porcentaje correspondiente que representan. Se puede observar cómo el número de términos que se utilizan para el aprendizaje se ve reducido considerablemente, con la consiguiente ventaja de la disminución del tiempo de aprendizaje.

3.4.2. Estimación de la información cuantitativa de la red mixta.

Para calcular las distribuciones de probabilidad marginales y también las condicionales, es decir, todas las distribuciones almacenadas en los nodos término, se utilizarán las mismas técnicas que las usadas para la redes simple y aumentada.

Con respecto a las distribuciones de probabilidad condicionadas correspondientes a los nodos documento, a las funciones de probabilidad ya diseñadas para la red aumentada vamos a añadir tres más que enmarcaremos dentro del enfoque basado en medidas de similitud, y que tienen por objeto determinar el impacto en la recuperación del hecho de incluir o no los términos que están aislados. Son las siguientes:

- *fp10c*: sólo intervienen en el cálculo de la probabilidad condicionada correspondiente los términos que están en el poliárbol, excluyendo a los términos aislados. De esta forma se simula el hecho de considerar que la red bayesiana documental estaría formada exclusivamente por el poliárbol de términos aprendido con los términos *conectados* y la subred de documentos. La forma de esta función de probabilidad es la siguiente:

$$p(d_j \mid \pi(D_j)) \propto \frac{\sum_{T_i \in D_j \cap C} tf_{ij} \cdot idf_i^2}{\sqrt{\sum_{T_i \in D_j \cap C} tf_{ij} \cdot idf_i^2}}$$

- *fp10d*: sólo intervienen los términos que pertenecen al poliárbol de la subred de términos más aquellos que, estando aislados, pertenecen a la consulta.

$$p(d_j \mid \pi(D_j)) \propto \frac{\sum_{T_i \in R_{\pi(D_j)} \cap C} tf_{ij} idf_i^2 + \sum_{T_i \in R_{\pi(D_j)} \cap \mathcal{A} \cap Q} tf_{ij} idf_i^2}{\sqrt{\sum_{T_i \in D_j \cap C} tf_{ij} idf_i^2 + \sum_{T_i \in R_{\pi(D_j)} \cap \mathcal{A} \cap Q} tf_{ij} idf_i^2}}$$

  Evaluar con esta función es análogo a tener una red “dinámica” que depende de la consulta: se parte de una red donde sólo figuran los términos incluidos en la clase C y se añaden a ella los términos aislados que figuran en la consulta que se haya suministrado al S.R.I. en cada momento, excluyendo a todos los demás. Por tanto, dependiendo de la consulta, se utilizarán unos u otros términos aislados, en contraposición a $fp10c$, donde están todos excluidos.

- *fp10e*: variante de $fp10d$ en la que se mantiene la idea de red “dinámica” en las mismas condiciones, pero con la diferencia fundamental de que la normalización que se lleva a cabo es diferente, interviniendo todos los términos del documento.

$$p(d_j \mid \pi(D_j)) \propto \frac{\sum_{T_i \in D_j \cap C} tf_{ij} idf_i^2 + \sum_{T_i \in R_{\pi(D_j)} \cap \mathcal{A} \cap Q} tf_{ij} idf_i^2}{\sqrt{\sum_{T_i \in D_j} tf_{ij} idf_i^2}}$$

3.5. El motor de recuperación: inferencia en la red bayesiana documental.

Una vez que tenemos construida una red bayesiana, ésta puede ser utilizada para realizar predicciones sobre los valores que puede tomar ciertas variables. A este proceso se le conoce como *inferencia* (cálculo de las probabilidades a posteriori). Los algoritmos que lo resuelven calculan las probabilidades de los distintos casos que puede tomar cada variable no conocida, dados los valores de las variables conocidas o evidencias.

Cuando nos centramos en la red bayesiana documental, la consulta (o más propiamente, los términos de la consulta) actúa como evidencia suministrada al sistema. Nuestro interés se centra en conocer las probabilidades de relevancia de las variables documento dada una consulta. Para ello, se instancian todos los términos que aparecen en la consulta a relevante. Esta información se propagará hacia los nodos documento para obtener $p(d_j \mid Q), \forall d_j$. Una vez conocidas estas probabilidades, los documentos se presentan al usuario ordenados en orden decreciente según el valor de $p(d_j \mid Q)$.

Independientemente del tipo de técnica usada para realizar la propagación, la eficiencia del proceso de propagación va a depender directamente del número de nodos que existan en la red y de la topología subyacente a ésta, es decir, el tipo de grafo. Así, si la complejidad del grafo es muy alta (el número de nodos de la red es grande), los procesos de propagación se vuelven muy costosos en tiempo.

Son dos los grupos de técnicas clásicas que resuelven el problema de inferencia en redes bayesianas: las exactas y las aproximadas [Pea88, CGH96, Sal98] . Las primeras obtienen las distribuciones a posteriori sobre las variables en las que estamos interesados mediante la aplicación de expresiones matemáticas exactas. Por otro lado, las segundas, basadas en métodos de simulación, persiguen el mismo objetivo que las primeras, consiguiéndolo en un tiempo más razonable, pero con el inconveniente de una pérdida de exactitud en los cálculos. Cuando aplicamos estas técnicas a nuestros modelos, encontramos los siguientes problemas:

- *Propagación con algoritmos exactos.*

  Teniendo en cuenta que la propagación es un problema NP-duro [Coo90] (no siempre se encuentran soluciones a un problema en tiempo polinomial), y considerando que:

  - en la topología propuesta aparecen ciclos,
  - el número elevado de padres de los nodos documento y
  - la gran cantidad de nodos existente,

  el tiempo consumido por un algoritmo de propagación exacto para realizar la inferencia en toda la red es prohibitivo, por lo que en este caso la propagación exacta está totalmente descartada.

- *Propagación con algoritmos aproximados.*

  Con los condicionantes expuestos anteriormente, la única solución es la utilización de un método de propagación aproximado [CGH96, Sal98]. Desde un punto de vista abstracto, el problema de la propagación aproximada se puede ver como la obtención de muestras a partir de una distribución de probabilidad difícil de manejar. Un tipo de estos algoritmos son los conocidos como *algoritmos de muestreo por importancia*, que utilizan una distribución modificada con objeto de obtener muestras independientes que son pesadas para parecerse a la original [Sal98].

Un inconveniente que tienen estos algoritmos es que comienzan la simulación por ciertas variables sin tener en cuenta la información almacenada en otras partes del grafo, problema que se acentúa cuando las redes son muy grandes y las probabilidades extremas. Dentro de los algoritmos de muestreo por importancia encontramos los basados en *precomputación aproximada*, los cuales llevan a cabo una primera fase en la que realizan una propagación no exacta, seguida de un proceso de eliminación de nodos. De esta forma, generan una distribución a posteriori. En la segunda fase, se obtiene una muestra a partir de esta distribución aproximada y estiman las probabilidades aplicando posteriormente una metodología de muestreo por importancia.

A su vez, clasificado como un algoritmo de este último tipo, Salmerón y col. [SCM00] han diseñado un método de propagación aproximado basado en árboles de probabilidad (el principal problema de los algoritmos de precomputación es que almacenan las distribuciones de probabilidad en tablas cuyo tamaño es proporcional al producto de los valores que puede tomar cada variable). Los árboles de probabilidad son grafos que permiten almacenar la información cuantitativa de manera más eficiente.

En este algoritmo, el proceso de eliminación de nodos al que nos referíamos anteriormente es especialmente apropiado para el caso en que haya pocas evidencias, ya que comienza borrando aquellas variables que no sean observadas y que no tengan descendientes también observados. Este hecho lo convierte en especialmente útil para propagar con redes bayesianas grandes. Además, si la secuencia de borrado es exacta, la distribución generada coincide plenamente con la buscada.

Dadas las características de nuestra red documental, teniendo en cuenta que los nodos documento no tienen descendientes y no se instancian como observados, el algoritmo calcula muestras exactas de la distribución, siendo la calidad del muestreo óptima. Estas características nos hicieron pensar en este algoritmo como idóneo para nuestros intereses.

La implementación que hemos utilizado para los experimentos en nuestro S.R.I. es la que figura dentro del software *Elvira* [Elv00], el cual es una herramienta genérica diseñada para trabajar con redes de creencia y que suministra al usuario un amplio abanico de algoritmos de aprendizaje y propagación en redes de creencia. Ahora bien, el problema con el que nos hemos encontrado es que la aplicación del algoritmo de propagación aproximada con las redes bayesianas documentales, consume un tiempo bastante considerable. Así, y a pesar de ser un algoritmo eficiente y particularmente bueno para la topología que tienen nuestras redes, su empleo no satisface unos mínimos requerimientos en cuanto al tiempo de cálculo. Por esta razón, sólo hemos podido probar el algoritmo de propagación con la colección más pequeña de las utilizadas, ADI.

Por tanto, ante la imposibilidad de aplicar cómodamente un algoritmo exacto de propagación clásico o un algoritmo aproximado, hemos de buscar soluciones que nos permitan realizar la inferencia en nuestros modelos en un tiempo aceptable.

3.5.1. Propagación + evaluación.

Esta técnica consiste en dividir el proceso de propagación en dos fases: una primera, en la que se realiza la propagación exacta en la subred de términos, y una segunda, en la que se utilizan las probabilidades a posteriori en los nodos término para evaluar las funciones de probabilidad almacenadas en los nodos documentos.

Así, teniendo en cuenta que las evidencias están reducidas a los nodos término que componen la consulta y que nunca se incluyen nodos documento, podemos utilizar el algoritmo de propagación exacta en poliárboles de Pearl [Pea88] para conocer la probabilidad a posteriori de cada nodo término, tal y como se hizo en el capítulo 2 para expandir la consulta. Estas probabilidades pueden calcularse en un tiempo polinomial de manera exacta.

Seguidamente, se procede a evaluar las funciones de probabilidad para cada documento, utilizando las probabilidades a posteriori estimadas en la fase anterior para modificar la fuerza con la que los distintos términos influyen en la relevancia de los documentos.

En este caso, ante una determinada consulta Q, la probabilidad de relevancia del documento se obtiene en función de las distintas probabilidades de relevancia del documento, dado cada uno de los términos individuales, $p(d_j \mid t_i), i = 1, \ldots, m_j$, ponderados por la probabilidad del término dada la consulta, $p(t_i \mid Q)$, es decir,

$$p(d_j \mid Q) = \bigotimes_{i=1}^{m_j} \big(p(d_j \mid t_i) \cdot p(t_i \mid Q)\big)$$

Veamos a continuación el siguiente teorema que es el que nos va a dar las condiciones bajo las cuales podremos realizar la propagación en dos fases con total equivalencia a la propagación exacta.

Teorema 3.1. *Dado un conjunto de evidencias correspondientes a los términos de una consulta Q, si la función de probabilidad utilizada se puede expresar cómo:*

$$p(d_j \mid \pi(D_j)) = \sum_{T_i \in R_{\pi(D_j)}} w_{ij}, \ \forall j = 1, \ldots, N \tag{3.8}$$

es decir, como una suma de pesos para sus términos relevantes, con $0 \leq w_{ij}, \forall i = 1, \ldots, m_j$ *y* $\sum_{T_i \in D_j} w_{ij} \leq 1$ *y siendo* $R_{\pi(D_j)}$ *el conjunto de términos que son relevantes en una configuración de padres de* D_j*,* $\pi(D_j)$*, entonces la propagación exacta en la subred de términos más la evaluación en la subred de documentos de las funciones de probabilidad en cada documento es equivalente a realizar una propagación exacta en la red bayesiana documental completa.*

Demostración.

La probabilidad a posteriori obtenida tras el proceso de inferencia exacta, $p(d_j \mid Q)$ se puede expresar como:

$$p(d_j \mid Q) = \sum_{\pi(D_j)} p(d_j \mid \pi(D_j), Q) \cdot p(\pi(D_j) \mid Q)$$

Como el conjunto de términos de un documento hace que éste y las evidencias sean independientes, entonces

$$p(d_j \mid Q) = \sum_{\pi(D_j)} p(d_j \mid \pi(D_j)) \cdot p(\pi(D_j) \mid Q)$$

Sustituyendo en la expresión anterior el valor de $p(d_j \mid \pi(D_j))$ de la ecuación (3.8) tenemos que

$$p(d_j \mid Q) = \sum_{\pi(D_j)} \left(\sum_{T_i \in R_{\pi(D_j)}} w_{ij} \cdot p(\pi(D_j) \mid Q) \right) \tag{3.9}$$

El siguiente paso será descomponer la sumatoria anterior en dos. En la primera incluimos las configuraciones donde el término T_{m_j} es relevante, y en la otra, donde no lo es.

$$\begin{array}{rcl} p(d_j \mid Q) & = & \sum_{\pi(D_j) \text{ y } T_{m_j} \in R_{\pi(D_j)}} \left(\sum_{T_i \in R_{\pi(D_j)}} w_{ij} \cdot p(\pi(D_j) \mid Q) \right) \quad + \\ & = & \sum_{\pi(D_j) \text{ y } T_{m_j} \notin R_{\pi(D_j)}} \left(\sum_{T_i \in R_{\pi(D_j)}} w_{ij} \cdot p(\pi(D_j) \mid Q) \right) \end{array} \tag{3.10}$$

Como

$$\sum_{T_i \in R_{\pi(D_j)}} w_{ij} \cdot p(\pi(D_j) \mid Q) = \sum_{T_i \in R_{\pi(D_j)} / T_i \neq T_{m_j}} w_{ij} \cdot p(\pi(D_j) \mid Q) + w_{m_j j} \cdot p(\pi(D_j) \mid Q)$$

Sustituyendo en el primer sumando de la expresión (3.10), la probabilidad a posteriori toma la forma:

$$\begin{array}{rcl} p(d_j \mid Q) & = & \sum_{\pi(D_j) \text{ y } T_{m_j} \in R_{\pi(D_j)}} \quad \Big(\sum_{T_i \in R_{\pi(D_j)} / T_i \neq T_{m_j}} w_{ij} \cdot p(\pi(D_j) \mid Q) + \\ & & \qquad + w_{m_j j} \cdot p(\pi(D_j) \mid Q) \Big) + \\ & + & \sum_{\pi(D_j) \text{ y } T_{m_j} \notin R_{\pi(D_j)}} \quad \left(\sum_{T_i \in R_{\pi(D_j)}} w_{ij} \cdot p(\pi(D_j) \mid Q) \right) \end{array}$$

Se observa cómo ambas sumas en las configuraciones, teniendo en cuenta y sin tenerlo a T_{m_j}, tienen en común $\sum_{T_i \in R_{\pi(D_j)} / T_i \neq T_{m_j}} w_{ij} \cdot p(\pi(D_j) \mid Q)$, por lo que mediante una operación de marginalización se podría unificar esos dos sumandos en uno único. En este caso, la suma correspondería a todas las configuraciones de los padres de D_j sin la variable T_{m_j}. Si hemos notado una configuración de los padres de D_j como $\pi(D_j) =< \mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_{m_j}} >$, introducimos ahora la nueva configuración $(\pi(D_j))^{\downarrow T_{m_j}} =< \mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_{m_j-1}} >$, donde se elimina la última variable, es decir, T_{m_j}.

Así, la probabilidad a posteriori del documento quedaría:

$$\begin{aligned} p(d_j \mid Q) &= \textstyle\sum_{(\pi(D_j))^{\downarrow T_{m_j}}} \sum_{T_i \in R_{(\pi(D_j))^{\downarrow T_{m_j}}}} w_{ij} \cdot p((\pi(D_j))^{\downarrow T_{m_j}} \mid Q) + \\ &+ \textstyle\sum_{\pi(D_j) \text{ y } T_{m_j} \in R_{\pi(D_j)}} w_{m_j j} p(\pi(D_j) \mid Q) \end{aligned}$$

Centrándonos en el segundo sumando de la igualdad anterior:

$$\sum_{\pi(D_j) \text{ y } T_{m_j} \in R_{\pi(D_j)}} w_{m_j j} p(\pi(D_j) \mid Q) = w_{m_j j} \cdot \sum_{\pi(D_j) \text{ y } T_{m_j} \in R_{\pi(D_j)}} p(\pi(D_j) \mid Q)$$

lo que implica que, al quedar una suma en todas las configuraciones, el resultado será $w_{m_j j} \cdot p(t_{m_j} \mid Q)$, valor que se obtiene a partir de la aplicación del proceso de propagación exacto en la subred de términos.

Así pues, tenemos

$$\begin{aligned} p(d_j \mid Q) &= \textstyle\sum_{(\pi(D_j))^{\downarrow T_{m_j}}} \sum_{T_i \in R_{(\pi(D_j))^{\downarrow T_{m_j}}}} w_{ij} \cdot p((\pi(D_j))^{\downarrow T_{m_j}} \mid Q) + \\ &+ w_{m_j j} \cdot p(t_{m_j} \mid Q) \end{aligned}$$

Nótese que el primer sumando es completamente análogo a la expresión de partida, ecuación (3.9), pero donde el término T_{m_j} ha sido eliminado. Seguidamente, reiteraríamos el proceso aplicado a este primer sumando, para eliminar una nueva variable T_{m_j-1} y extraer el sumando $w_{m_j-1j} \cdot p(t_{m_j} \mid Q$. Continuando de esta forma hasta eliminar todos los términos, se llega a la expresión final de la probabilidad a posteriori de un documento dadas las evidencias:

$$p(d_j \mid Q) = \sum_{i=1}^{m_j} w_{ij} \cdot p(t_i \mid Q)$$

Así pues, podemos calcular de forma exacta la probabilidad $p(d_j \mid Q)$ realizando únicamente propagación exacta en la subred de términos.

$\square$

Las funciones de probabilidad que cumplen las condiciones expresadas por el teorema anterior, para un documento D_j, son las mostradas en la tabla 3.2, donde c es una constante de normalización, para garantizar que la suma de los pesos w_{ij} es menor o igual que 1 y que, por tanto, los valores devueltos por la función de probabilidad son realmente probabilidades. Obsérvese que al tratarse de una constante global, la ordenación de los documentos es la misma incluyendo c en los cálculos o no.

Denominación	w_{ij}
fp2	$\frac{tf_{ij} \cdot idf_i}{\sum_{l=1}^{m_j} tf_{lj}^2 \cdot idf_l^2}$
fp3	$\frac{1}{m_j}$
fp4	$\frac{tf_{ij}^2 \cdot idf_i^2}{\sum_{l=1}^{m_j} tf_{lj}^2 \cdot idf_l^2}$
fp6	$\frac{tf_{ij}^2 \cdot idf_i^2}{c \cdot \sqrt{\sum_{l=1}^{m_j} tf_{lj}^2 \cdot idf_l^2}}$
fp8	$\frac{tf_{ij} \cdot idf_i^2}{\sum_{l=1}^{m_j} tf_{lj} \cdot idf_l^2}$
fp10	$\frac{tf_{ij} \cdot idf_i^2}{c \cdot \sqrt{\sum_{l=1}^{m_j} tf_{lj} \cdot idf_l^2}}$
fp10c	$\frac{tf_{ij} \cdot idf_i^2 \cdot I_C(T_i)}{c \cdot \sqrt{\sum_{T_i \in D_j \cap C} tf_{ij} \cdot idf_i^2}}$
fp10d	$\frac{tf_{ij} \cdot idf_i^2 \cdot I_{C \cup (\mathcal{A} \cap Q)}(T_i)}{c \cdot \sqrt{\sum_{T_i \in D_j \cap C} tf_{ji} \cdot idf_i^2 + \sum_{T_i \in D_j \cap \mathcal{A} \cap Q} tf_{ij} \cdot idf_i^2}}$
fp10e	$\frac{tf_{ij} \cdot idf_i^2 \cdot I_{C \cup (\mathcal{A} \cap Q)}(T_i)}{c \cdot \sqrt{\sum_{T_i \in D_j} tf_{ij} \cdot idf_i^2}}$

Cuadro 3.2.: Funciones de probabilidad que cumplen la condición del teorema 3.1.

Denominación	Expresión
$fp5$	$p(d_j \mid Q) = 1 - \Pi_{T_i \in D_j}(1 - p(d_j \mid t_i) \cdot p(t_i \mid Q))$
$fp1$	$p(d_j \mid Q) = 1 - \Pi_{T_i \in D_j}(1 - p(d_j \mid t_i) \cdot p(t_i \mid Q) + p(d_j \mid \bar{t}_i) \cdot p(\bar{t}_i \mid Q))$

Cuadro 3.3.: Funciones de probabilidad que no cumplen las condiciones del teorema 3.1.

El símbolo $I_B(T_i)$ usado en la tabla representa la función indicadora de un conjunto, es decir, $I_B(T_i) = 1$ si $T_i \in B$ e $I_B(T_i) = 0$ si $T_i \notin B$.

El resto de funciones de probabilidad también se pueden utilizar para realizar la evaluación de los documentos.Aunque, en este caso no cumplen la propiedad que hace que su uso origine los mismos resultados que se conseguirían efectuando la propagación exacta. Estas funciones de probabilidad quedan como se muestra en la tabla 3.3. Sin embargo, cuando nos encontramos en la red simple podemos garantizar que la evaluación de la función *fp5* también proporciona resultados exactos.

Con el siguiente teorema demostraremos que la propagación en la red simple con la función de probabilidad *fp5* es exacta.

Teorema 3.2. *La propagación en la red simple mediante la evaluación de la función de probabilidad fp5 es exacta, esto es,*

$$p(d_j \mid Q) = 1 - \prod_{i=1}^{m_j}(1 - p(t_i \mid Q) \cdot p(d_j \mid t_i))$$

Demostración.

Dado un documento $D_j = (T_1, \cdots, T_k, T_{k+1}, \cdots, T_{m_j})$, y una consulta Q, ésta podemos dividirla en dos conjuntos de términos disjuntos: el de los que pertenecen al documento, $Q_{D_j} = (t_1, \cdots, t_k)$, y el de los términos que no indexan a D_j.

Para realizar la demostración, consideraremos que en la red simple, se cumple que $p(\bar{d}_j \mid Q) = p(\bar{d}_j \mid Q_{D_j})$.

Así,

$$p(\bar{d}_j \mid Q_{D_j}) = \sum_{t_{k+1}, \cdots, t_{m_j}} p(\bar{d}_j \mid t_{k+1}, \cdots, t_{m_j}, Q_{D_j}) \cdot p(t_{k+1}, \cdots, t_{m_j} \mid Q_{D_j})$$

Como los términos son independientes entre sí en este modelo, tenemos que:

$$p(\bar{d}_j \mid Q_{D_j}) = \sum_{t_{k+1},\cdots,t_{m_j}} p(\bar{d}_j \mid t_{k+1},\cdots,t_{m_j},Q_{D_j}) \cdot \prod_{i=k+1}^{m_j} p(t_i \mid Q_{D_j})$$

Considerando la expresión *fp5* tenemos que:

$$p(\bar{d}_j \mid t_{k+1},\cdots,t_{m_j},Q_{D_j}) = 1 - p(d_j \mid t_{k+1},\cdots,t_{m_j},Q_{D_j}) = \prod_{i=1}^{m_j}(1-p(d_j \mid t_i))$$

Por tanto, $p(\bar{d}_j \mid Q_{D_j})$ se puede expresar como:

$$p(\bar{d}_j \mid Q_{D_j}) = \sum_{t_{k+1},\cdots,t_{m_j}} \prod_{i=1}^{m_j}(1-p(d_j \mid t_i)) \cdot \prod_{i=k+1}^{m_j} p(t_i \mid Q_{D_j})$$

Saquemos a continuación el término T_{m_j} de la sumatoria:
$p(\bar{d}_j \mid Q_{D_j}) =$

$$\begin{aligned} = \quad & p(t_{m_j} \mid Q_{D_j}) \cdot \textstyle\sum_{t_{k+1},\cdots,t_{m_j-1}} \left[\prod_{i=1}^{m_j-1}(1-p(d_j \mid t_i))\right](1-p(d_j \mid t_{m_j})\prod_{i=k+1}^{m_j-1} p(t_i \mid Q_{D_j}) + \\ + \quad & p(\bar{t}_{m_j} \mid Q_{D_j}) \cdot \textstyle\sum_{t_{k+1},\cdots,t_{m_j-1}} \left[\prod_{i=1}^{m_j-1}(1-p(d_j \mid t_i))\right](1-p(d_j \mid \bar{t}_{m_j})\prod_{i=k+1}^{m_j-1} p(t_i \mid Q_{D_j}) \end{aligned}$$

Como $p(\bar{d}_j \mid t_{m_j}) = 0$ y $p(\bar{t}_{m_j} \mid Q_{D_j}) = 1 - p(t_{m_j} \mid Q_{D_j})$, entonces tenemos que
$p(\bar{d}_j \mid Q_{D_j}) =$

$$\begin{aligned} = \quad & p(t_{m_j} \mid Q_{D_j}) \textstyle\sum_{t_{k+1},\cdots,t_{m_j-1}} \left[\prod_{i=1}^{m_j-1}(1-p(d_j \mid t_i))\right](1-p(d_j \mid t_{m_j})\prod_{i=k+1}^{m_j-1} p(t_i \mid Q_{D_j}) + \\ + \quad & (1-p(t_{m_j} \mid Q_{D_j})) \cdot \textstyle\sum_{t_{k+1},\cdots,t_{m_j-1}} \prod_{i=1}^{m_j-1}(1-p(d_j \mid t_i))\prod_{i=k+1}^{m_j-1} p(t_i \mid Q_{D_j}) \end{aligned}$$

Sacando la sumatoria como factor común y operando:

$$\begin{aligned} p(\bar{d}_j \mid Q_{D_j}) \quad &= \quad \left[p(t_{m_j} \mid Q_{D_j})(1-p(d_j \mid t_{m_j})) + (1-p(t_{m_j} \mid Q_{D_j}))\right] \cdot \\ & \quad\quad \textstyle\sum_{t_{k+1},\cdots,t_{m_j-1}} \prod_{i=1}^{m_j-1}(1-p(d_j \mid t_i))\prod_{i=k+1}^{m_j-1} p(t_i \mid Q_{D_j}) = \\ &= \quad (1-p(t_{m_j} \mid Q_{D_j}) \cdot p(d_j \mid t_{m_j})) \cdot \\ & \quad\quad \textstyle\sum_{t_{k+1},\cdots,t_{m_j-1}} \prod_{i=1}^{m_j-1}(1-p(d_j \mid t_i))\prod_{i=k+1}^{m_j-1} p(t_i \mid Q_{D_j}) = \end{aligned}$$

Repitiendo el mismo proceso para todos los términos que no están en la consulta, llegamos a la siguiente expresión:

$$p(\bar{d}_j \mid Q_{D_j}) = \prod_{i=k+1}^{m_j}(1-p(t_{m_j} \mid Q_{D_j}) \cdot p(d_j \mid t_{m_j}))\prod_{i=1}^{k}(1-p(d_j \mid t_i))$$

Si tenemos en cuenta que $p(t_i \mid Q) = 1$ si $T_i \in Q$, la expresión anterior se puede poner como:

$$p(\bar{d}_j \mid Q) = \prod_{i=1}^{m_j}(1 - p(t_i \mid Q) \cdot p(d_j \mid t_i)), \text{ por lo que } p(d_j \mid Q) = 1 - \prod_{i=1}^{m_j}(1 - p(t_i \mid Q) \cdot p(d_j \mid t_i))$$

□

Por tanto, cuando trabajamos con la red simple la propagación es exacta con *fp5*. En el caso de tomar la red aumentada o mixta, el proceso de propagación + evaluación se puede considerar equivalente a instanciar los distintos términos de una red simple con las probabilidades a posteriori del término dada la consulta que ha sido instanciada en el poliárbol de términos.

3.5.2. Métodos para incluir la importancia de los términos en el proceso de inferencia.

A lo largo de todo el capítulo hemos asumido que la presencia de un término en la consulta implica su instanciación a relevante. Con este criterio, se le está asignando la misma "fuerza" a todos los términos de la consulta. Sin embargo, muchas veces, al efectuar una pregunta al S.R.I., podemos estar interesados en destacar la importancia de un término frente a otros que pueden ser secundarios. Este tipo de información se incorpora de forma natural en el modelo vectorial [SM83].

En esta sección vamos a presentar un par de métodos desarrollados con el objetivo fundamental de incluir en el proceso de propagación la importancia propia de cada término. Con ello, pretendemos mejorar el rendimiento recuperador ofrecido por el modelo expuesto en este capítulo.

Las técnicas propuestas permiten incluir dos componentes distintos de información sobre el término, ampliando la versatilidad del modelo desarrollado: por un lado, basándonos en el concepto de *evidencia parcial*, se pretende que la probabilidad a posteriori de un término de la consulta dependa de su calidad como término; por otro, se intenta que los términos que aparecen más veces en la consulta tengan una mayor influencia a la hora de propagar que los que figuran pocas veces.

Ambos métodos no son excluyentes entre sí en cuanto a su uso, es decir, son técnicas totalmente complementarias, cuyo uso simultáneo puede dar lugar a un incremento de la capacidad recuperadora del S.R.I.

3.5.2.1. Instanciación parcial de evidencias.

Nuestro objetivo será permitir que un término de la consulta pueda ser considerado como parcialmente relevante, es decir, admitiremos que tenga valores de probabilidad a posteriori menores que uno. En este caso, la probabilidad a posteriori del término instanciado dependerá de su calidad como término y podrá ser distinta de la del resto de términos de la consulta, así como variar entre las distintas consultas. Para conseguir este objetivo, utilizaremos el concepto de *evidencia parcial*.

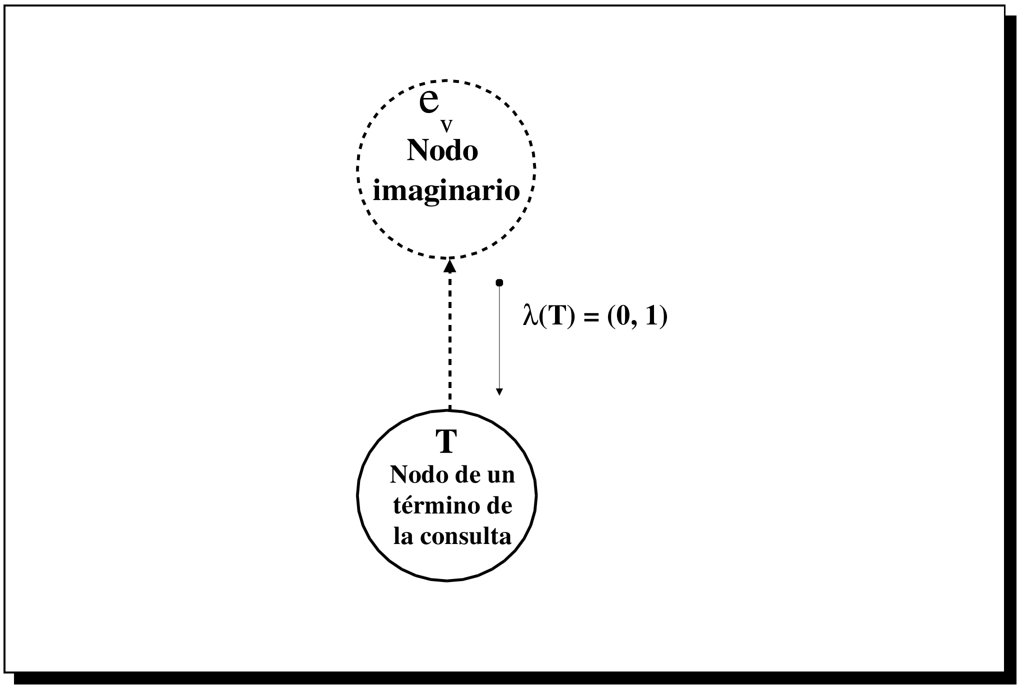


Figura 3.8.: Instanciación de un nodo término de la red bayesiana documental.

La instanciación de una variable como observada se podría ver como la creación de un nodo imaginario, llamémosle *ev*, que se sitúa como hijo del nodo término observado en la consulta. Este nodo virtual, cuando comience la inferencia, enviará al nodo instanciado un vector $\lambda(T_i) = (\lambda(\bar{t}_i), \lambda(t_i))$, como se puede ver en la figura 3.8. Cada uno de los componentes del vector λ son verosimilitudes y representan la probabilidad de que las evidencias (observaciones) sean relevantes dado que el término es no relevante y relevante, respectivamente. De forma general,

$$\lambda(T_i) = (P(\ Observación\ \mid \bar{t}_i), p(\ Observación\ \mid t_i)),$$

Una vez que el nodo término recibe el vector λ, lo combina con la probabilidad a priori para calcular $p(T_i \mid ev)$. Para ello, aplicando el teorema de Bayes se obtiene que:

$$p(T_i \mid ev) = a \cdot p(ev \mid T_i) \cdot p(T_i),$$

siendo a una constante de normalización. Utilizando una notación vectorial, esta expresión anterior es equivalente a:

$$\begin{aligned} p(T_i \mid ev) &= a \cdot (p(ev \mid \bar{t}_i), p(ev \mid t_i)) \cdot (p(\bar{t}_i), p(t_i)) = \\ &= a \cdot (\lambda(\bar{t}_i), \lambda(t_i)) \cdot (p(\bar{t}_i), p(t_i)) = \\ &= a \cdot (\lambda(\bar{t}_i) \cdot p(\bar{t}_i), \lambda(t_i) \cdot p(t_i)) \end{aligned}$$

Si llamamos $p(t_i) = p_i$ y $p(\bar{t}_i) = 1 - p_i$, sustituyendo en la expresión anterior y normalizando por la suma de los componentes del vector, la probabilidad a posteriori queda:

$$p(T_i \mid ev) = (p(\bar{t}_i \mid ev), p(t_i \mid ev)) = \left(\frac{(1-p_i)\lambda(\bar{t}_i)}{p_i\lambda(t_i) + (1-p_i)\lambda(\bar{t}_i)}, \frac{p_i\lambda(t_i)}{p_i\lambda(t_i) + (1-p_i)\lambda(\bar{t}_i)} \right) \tag{3.11}$$

En nuestro caso, al instanciar el término a relevante, el vector λ enviado desde el nodo imaginario será $(0,1)$. Con estos valores, nos aseguramos que $p(t_i \mid ev) = 1$.

Nuestra intención ahora será que el término T_i posea una probabilidad a posteriori dependiente de la calidad del término. La pregunta que se nos plantea inmediatamente es qué vector λ debe mandarle el nodo imaginario para poder tener como probabilidad a posteriori final un valor q_i, esto es, $p(t_i \mid evidencia) = q_i$. Esta probabilidad que se fija para un nodo T_i es la que tomaría ese nodo en ausencia de cualquier otra evidencia.

Una propiedad interesante a considerar relacionada con los vectores λ es que, en realidad, no interesan las magnitudes de los dos elementos, sino la proporción que hay entre ellos. Por tanto, podemos fijar $\lambda(t_i) = 1$ y calcular el valor de $\lambda(\bar{t}_i)$. Como

$$p(t_i \mid ev) = \frac{p_i}{p_i + (1-p_i)\lambda(\bar{t}_i)} = q_i$$

Despejando p_i tendremos que:

$$p_i = p_i q_i + (1-p_i)\lambda(\bar{t}_i) q_i$$

La verosimilitud cuando el término es no relevante, volviendo a despejar esta vez $\lambda(\bar{t}_i)$, será:

$$\lambda(\bar{t}_i) = \frac{(1-q_i)p_i}{(1-p_i)q_i}$$

Por último, debemos garantizar que $0 \leq \lambda(\bar{t}_i) \leq 1$, ya que si $\lambda(\bar{t}_i) > 1$, estaríamos favoreciendo la no relevancia frente a la relevancia del término, lo que no parece sensato si tenemos en cuenta que ti pertenece a la consulta.

- Que $\lambda(\bar{t}_i) \geq 0$, es trivial.

- Veamos que $\lambda(\bar{t}_i) \leq 1$:

  $$\frac{(1-q_i)p_i}{(1-p_i)q_i} \leq 1 \Leftrightarrow (1-q_i)p_i \leq (1-p_i)q_i \Leftrightarrow p_i - p_iq_i \leq q_i - p_iq_i \Leftrightarrow p_i \leq q_i$$

  Lo cual no debe dar problemas pues parece lógico pensar que, para un término que pertenece a la consulta, la probabilidad a priori de relevancia del mismo será menor que su probabilidad a posteriori.

Por tanto, el vector λ que manda el nodo imaginario al nodo instanciado, como podemos ver en la figura 3.9, es:

$$\lambda(T_i) = \left(\frac{(1-q_i)p_i}{(1-p_i)q_i}, 1\right)$$

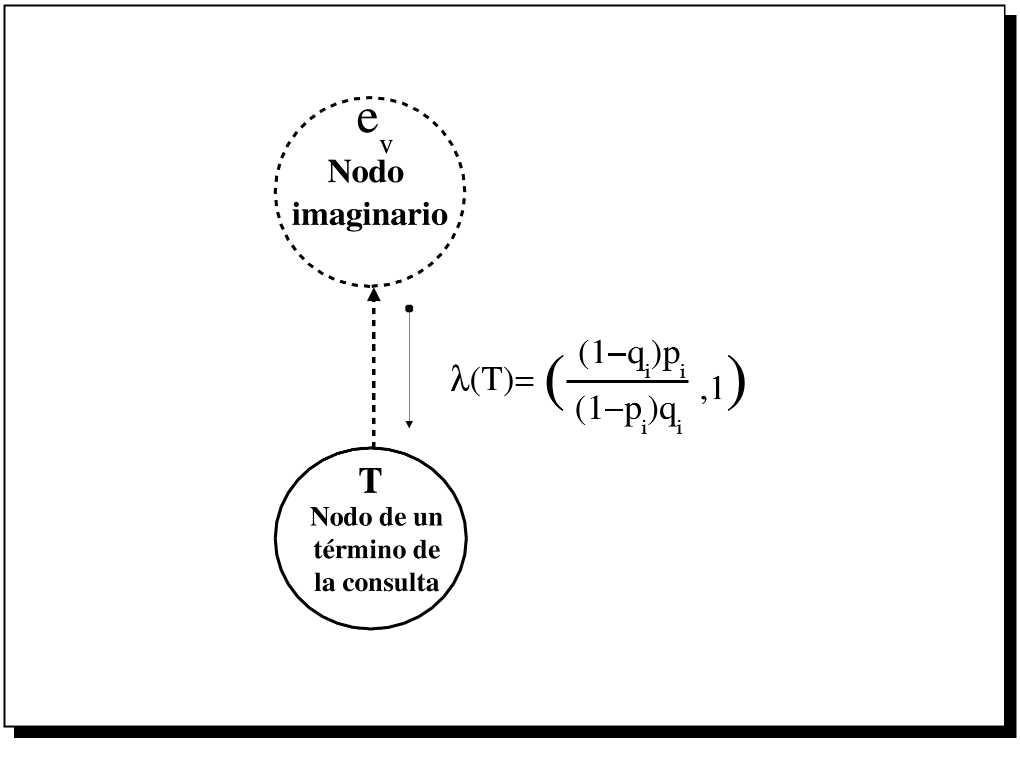


Figura 3.9.: Instanciación parcial de un nodo término de la red bayesiana documental.

Finalmente, nos queda presentar la propuesta de la probabilidad q_i para el término T_i, dado que éste está en la consulta. Para ello, utilizaremos una expresión con la que trabaja también Turtle [Tur90], aunque este autor la usa en un contexto distinto.

$$p(t_i \mid T_i \in Q) = \alpha + (1-\alpha)\frac{tf_{iQ} \cdot idf_i}{max_{T_j \in Q} tf_{jQ} \cdot idf_j} \tag{3.12}$$

siendo $0 \leq \alpha \leq 1$. Esta expresión corresponde a una combinación convexa del producto del $tf \cdot idf$ del término en cuestión, normalizado por el máximo de estos pesos de los términos que aparecen en la consulta. Así, cuando $\alpha = 0$, la probabilidad a posteriori coincidirá con el cociente anterior, al contrario que ocurre cuando $\alpha = 1$, momento en el cual la probabilidad a posteriori pasará directamente a ser uno.

3.5.2.2. Inclusión de la frecuencia de los términos de la consulta.

La idea básica de esta técnica es introducir en la evaluación de la función de probabilidad de los documentos la frecuencia de aparición de cada término en la consulta (si aparece), es decir, su qf.

En nuestro modelo, la inclusión del qf se simula mediante la replicación de los nodos término que pertenecen a la consulta. Así, si la frecuencia de aparición de un término T_i en la consulta es de tres ocurrencias, se crearían dos nodos ficticios, con la misma información que contiene el nodo T_i, y serían utilizados en la evaluación de cualquier documento que contuviera T_i. De esta manera, este nodo estaría incluido tres veces en la red bayesiana y contaría en el peso de relevancia de los documentos que lo contengan tantas veces como valor tuviera su qf_i.

Para facilitar, por tanto, el poder introducir la importancia relativa de los términos en la consulta, se ha pensado en incluir la frecuencia del término en ella en las funciones de probabilidad.

Así, por ejemplo, la probabilidad de relevancia de un documento D_j dada la consulta, utilizando la función $fp10$, quedaría como sigue:

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

En esta expresión se puede ver cómo se ha añadido en las sumas del numerador y del denominador el factor $[qf_i]$ que tomará el valor 1 si el i-ésimo término no está en la consulta y el correspondiente qf_i si lo está, razón por la cual se ha notado entre corchetes [3].

Esta forma de premiar los términos atendiendo al número de apariciones en la consulta es equivalente a la replicación del término instanciado, tantas veces como valor tenga su qf. Por ejemplo, consideremos cómo quedaría conceptualmente la red documental cuando se realiza la siguiente consulta: $((T_1, 1), (T_4, 2), (T_{10}, 3))$.

En la figura 3.10, los nodos con trazo discontinuo y fondo blanco son los nodos evidencias. Los nodos T_4 y T_{10}, al tener un qf distinto de uno, se duplican y triplican, respectivamente. De esta forma, al aplicar la función de relevancia correspondiente a los documentos que contengan esos términos aportarán el doble y el triple en peso al peso total devuelto por la función.

[3] Más formalmente, podríamos definir el factor $[qf_i]$ de la siguiente forma: $[qf_i] = 1 + (qf_i - 1) \cdot I_Q(T_i)$, donde $I_Q(Ti)$ es la función indicadora del conjunto Q de términos de la consulta.

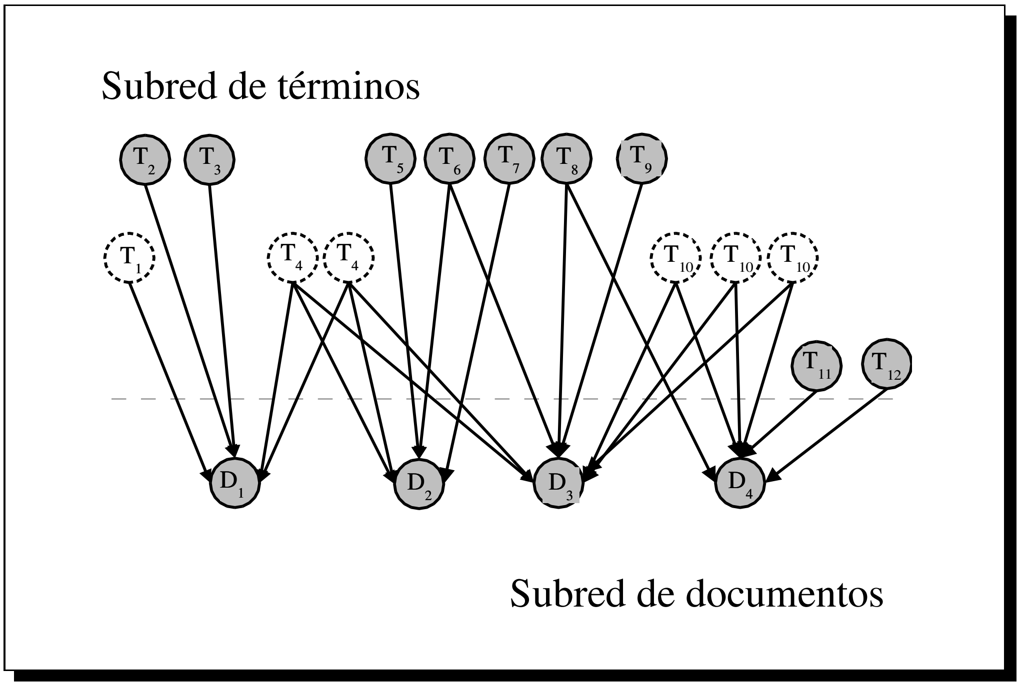


Figura 3.10.: Replicación de los nodos relacionados con los términos de una consulta.

Aunque en el ejemplo de la figura 3.10 se haya utilizado una red simple, esta técnica se puede poner igualmente en práctica con los otros dos tipos de redes. Así, una vez replicado un término T_i cualquiera originando, por ejemplo, dos nuevos nodos, T'_i y T''_i, éste se conecta con sus réplicas mediante un arco que los une con ellas. Por tanto, los nodo réplica sólo estarían enlazados a T_i y a todos los nodos documento indexados por T_i, como se puede ver en la figura 3.11. Además los valores de probabilidad son una copia idéntica de los que almacena el padre, para lo cual los nodos réplica almacenarán la siguiente distribución de probabilidad (en este caso T_i'):

$$\begin{array}{ll} p(\bar{t'}_i \mid \bar{t}_i) = 1 & p(\bar{t'}_i \mid t_i) = 0 \\ p(t'_i \mid \bar{t}_i) = 0 & p(t'_i \mid t_i) = 1 \end{array}$$

Con este tipo de matriz se evita que el proceso de propagación se vea afectado por la incorporación de estos nuevos nodos.

Comparativamente hablando, podríamos concluir que este método general permite premiar a los términos de la consulta, mientras que el anterior, el basado en la instanciación parcial, tendría el efecto contrario, es decir, penalizaría a los términos peores, mientras que a los mejores, como máximo los dejaría en una instanciación total a relevante.

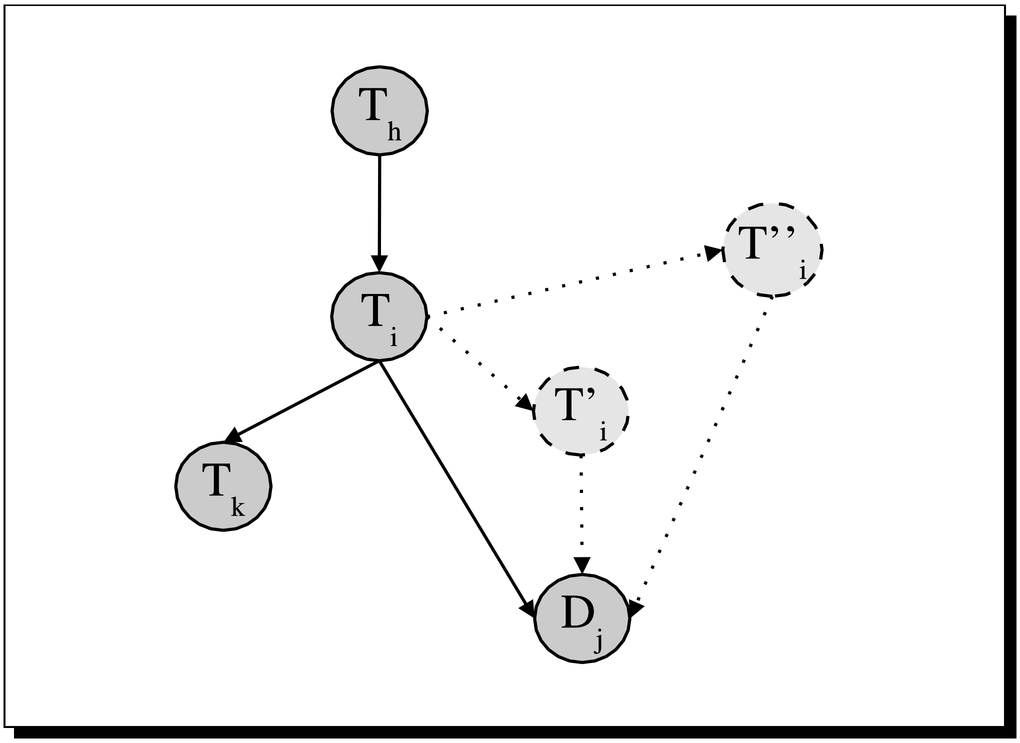


Figura 3.11.: Replicación de un nodo en un grafo cualquiera.

3.5.3. Ordenación de los documentos según la diferencia de probabilidades a posteriori y a priori de éstos.

El hecho de que la probabilidad a posteriori de un documento sea alta después del proceso de propagación, se puede deber a dos motivos fundamentales:

- Que el documento se vea positivamente influenciado por la instanciación de los términos de la consulta, o
- que su probabilidad a priori sea alta y la influencia que recibe de la consulta no decremente la creencia en la relevancia del documento.

El primer caso es el que se intenta detectar ya que refleja que el documento es bastante relevante con respecto a la consulta. Sin embargo, en el segundo caso, al considerar los documentos ordenados según el valor de la probabilidad a posteriori, podemos cometer algún error al dar una mayor importancia a un documento que apenas ha incrementado su creencia, con el consiguiente ruido que aporta a la recuperación.

Consideremos el siguiente ejemplo. Supongamos que las probabilidades a priori de tres documentos son las siguientes: $p(d_1) = 0{,}2, p(d_2) = 0{,}1, p(d_3) = 0{,}1$, y que tras propagar, las probabilidades a posteriori obtenidas son: $p(d_1 \mid Q) = 0{,}21, p(d_2 \mid Q) = 0{,}18, p(d_3 \mid Q) = 0{,}15$.

Al considerar sólo las probabilidades y elaborar la ordenación según ellas, el más relevante sería D_1, seguido por D_2 y, finalmente, por D_3. Pero, si nos fijamos en la probabilidad a priori del primer documento, observamos que es prácticamente la misma que su a posteriori, indicando por tanto que no ha recibido mucha influencia de la consulta al propagar. Esto implica que no es realmente el más relevante de los tres. Si nos fijamos en el incremento de las probabilidades más que en las magnitudes en sí, la ordenación quedaría ahora D_2, D_3, D_1.

Por tanto, la elaboración de la ordenación de documentos se podría hacer también según la diferencia $p(d_j \mid Q) - p(d_j), \forall D_j$.

Esta nueva ordenación de documentos puede verse como una nueva función de probabilidad, donde la probabilidad a posteriori del documento, $p'(d_j \mid Q)$ se obtiene ponderando el peso de cada término por la diferencia entre probabilidades a posteriori y a priori de cada término, $p(t_i \mid Q) - p(t_i)$. Con ello, lo que realmente influye en el valor de relevancia final del documento es la cuantía del incremento y no el valor relativo de la probabilidad.

Cualquier función de probabilidad mostrada anteriormente puede ser adaptada para incorporar esta nueva peculiaridad. En el caso de $fp10$, la probabilidad del documento dada la consulta se obtendrá como sigue a continuación.

$$p(d_j \mid Q) = \frac{\sum_{T_i \in D_j} tf_{ij} idf_i^2 (p(t_i \mid Q) - p(t_i))[qf_i]}{\sqrt{\sum_{T_i \in D_j} tf_{ij} idf_i^2 [qf_i]}}$$

Para notar que la función de probabilidad está utilizando la diferencia de probabilidades añadiremos un signo menos al final de cada uno de los nombres de las funciones. Así, para el ejemplo anterior, la nueva función pasaría a notarse como $fp10-$.

Es interesante hacer ver que, en la red simple el hecho de tener en cuenta la diferencia de probabilidades supone que sólo se evalúen los términos que pertenecen a la consulta, no considerando los que no pertenecen, ya que para éstos últimos, las probabilidades a posteriori son iguales que las a priori.

3.6. Experimentación con el modelo de recuperación de la red bayesiana documental.

En esta sección vamos a exponer la experimentación que hemos realizado para determinar la calidad del modelo de red bayesiana documental con cada una de las cinco colecciones estándar de prueba. La experimentación se ha organizado según el tipo de red bayesiana que hemos descrito en este capítulo.

En cada caso, nuestro objetivo es mostrar cuál es el comportamiento de los modelos propuestos. Para ello nos centraremos en estudiar los aspectos básicos. Por un lado, cómo se comportan los distintos modelos cuando se consideran diferentes métodos para estimar las distribuciones de probabilidad marginales y condicionadas almacenadas en la red. Por otro, se pretende evaluar el rendimiento con los métodos alternativos a la hora de realizar la propagación.

Con objeto de evitar la explosión combinatoria que supondría probar cada modelo con todos las posibles alternativas, se han planificado las distintas pruebas por etapas. En cada una de ellas se ha intentado determinar el impacto de un conjunto de valores para algunos parámetros en concreto, fijando el resto a un valor dado. Del estudio de los resultados obtenidos, se selecciona el mejor valor o conjunto de valores para los parámetros estudiados, que serán los que se queden fijos para fases posteriores.

Para mostrar la calidad de los resultados conseguidos por nuestros modelos, los vamos a comparar con los obtenidos por el S.R.I. *SMART*, ya que es una aplicación que está disponible como software de libre distribución y ha sido ampliamente utilizado como punto de comparación de la calidad recuperadora de multitud de sistemas. El esquema de ponderación con que hemos efectuado las recuperaciones con *SMART* es "ntc", debido a que es uno con los que este S.R.I. alcanza mejores resultados con las colecciones de prueba que hemos utilizado.

Para cada experimento, ofrecemos el número de documentos relevantes recuperados (REL. REC.), la precisión media para los once puntos de exhaustividad (M. 11PTS) y el porcentaje de cambio que se alcanza con respecto a la media de los once puntos conseguida por *SMART*. Los resultados completos, incluidas las curvas exhaustividad - precisión figuran en el anexo B. Al final de este capítulo compararemos los resultados de nuestro modelo con aquellos disponibles obtenidos por otros S.R.I. basados en redes bayesianas.

Para presentar la experimentación que hemos llevado a cabo, comenzaremos con el modelo simple, para pasar seguidamente al aumentado y concluir con la red mixta, aunque antes determinaremos cuál será el mejor método para realizar la inferencia en estos modelos. Aunque cada batería de experimentos se acompaña de los correspondientes comentarios y conclusiones, al final de la experimentación (sección 3.6.5) se exponen, como resumen, las principales conclusiones obtenidas.

3.6.1. Elección del método de propagación.

En primer lugar, y antes de continuar, hay que responder a la pregunta: "¿Qué método de propagación escoger para realizar los experimentos?" Como ya hemos dicho anteriormente, disponemos de tres tipos: propagación exacta (descartada directamente por su costo), propagación aproximada (utilizando el algoritmo de muestreo por importancia de Salmerón y col. [SCM00]) o propagación exacta en la subred de términos más evaluación de las funciones de probabilidad en la subred de documentos.

En un experimento inicial, efectuamos la recuperación con la red simple tanto con el algoritmo de propagación aproximada como con la propagación exacta más la evaluación. En este experimento consideramos los estimadores de distribuciones marginales *pp1, pp2* y *pp3*, y para las funciones de probabilidad *fp1* y *fp3*. Los resultados se muestran en la tabla 3.4.

Exp.	SMART	fp1, pp1	fp1, pp2	fp1, pp3	fp3, pp1	fp3, pp2	fp3, pp3
P. Aprox.							
REL. REC.	91	34	87	87	69	82	79
M. 11PTS	0.4706	0.1311	0.3870	0.3508	0.2729	0.3192	0.2770
%C. 11PTS		-72.14	-17.8	-23.9	-42.0	-32.2	-41.13
P. E. + E.							
REL. REC.	91	26	25	25	82	82	82
M. 11PTS	0.4706	0.1144	0.1130	0.1131	0.3098	0.3098	0.3098
%C. 11PTS		-75.7	-76.0	-76.0	-34.2	-34.2	-34.2

Cuadro 3.4.: Batería I con la red simple de ADI: propagación aproximada y exacta + evaluación.

Observando los resultados de esta tabla, podríamos decir que el uso de la propagación aproximada ofrece muchas veces mejores resultados que la propagación exacta más la evaluación. Sin embargo, son dos los problemas fundamentales que tiene la primera frente a la segunda: el primer es que es mucho más costosa, desde el punto de vista del tiempo de ejecución. Al ejecutarse sobre un software de propósito general, como es *Elvira*, éste no está optimizado para tratar con redes de un tamaño considerable, típicas en este campo de la R.I. Es por esto que se requiere el desarrollo de soluciones a medida que alivien este problema. Estas soluciones pueden venir de la mano de algoritmos de propagación totalmente adaptados a las características de las redes de que disponemos. Este es el caso de la propagación exacta más la evaluación, donde aprovechando las características topológicas de las redes con las que trabajamos evitamos realizar propagaciones en la red completa: se sustituyen por propagaciones exactas en una parte de ella y evaluaciones de funciones de probabilidad en la otra. Esta forma de proceder aporta la ventaja adicional de que se aseguran los mismos resultados que poniendo en práctica el proceso de inferencia exacta en toda la red [4] y se reduce considerablemente el tiempo de ejecución.

Además, obsérvese cómo los experimentos en que la propagación aproximada da mejores resultados se refieren esencialmente al caso en que se emplea *fp1*, con lo que puede estar ocurriendo simplemente que unos valores de probabilidad, aunque sean aproximados, produzcan mejores resultados que los valores calculados por la propagación exacta más la evaluación de la fórmula *fp1*, que no son sino otra aproximación peor comportada de las probabilidades exactas. En cambio, la tendencia se invierte al emplear *fp3*, que sí garantiza resultados exactos. Todo indica, en la propagación aproximada a que la probabilidad de relevancia de los documentos se muestra bastante sensible al muestreo realizado, el cual, dado el gran número de padres de un documento parece ser poco fiable.

Por estas razones, elegiremos la propagación exacta + evaluación como método de inferencia para el resto de la experimentación.

[4] Siempre que se empleen funciones de probabilidad que satisfagan las condiciones del teorema 3.1.

3.6.2. Experimentación con la red simple

En la red simple, disponemos de los siguientes parámetros de prueba, agrupados en clases, con sus correspondientes valores:

1. *Estimación de distribuciones de probabilidad:*
   - Estimadores de distribuciones de probabilidad marginales: *pp1, pp2, pp3* y *pp4*.
2. *Propagación:*
   - Funciones de probabilidad: *fp1, fp2, fp3, fp4, fp5, fp6, fp8* y *fp10*.
   - Intervención de la frecuencia del término en la consulta (*con qf*) y no intervención (*sin qf*).
   - Uso de la diferencia entre las probabilidades a posteriori y a priori del documento o sólo la probabilidad a posteriori.

3.6.2.1. Batería de experimentos I.

Para resolver estos experimentos hemos utilizado el método de propagación exacta + evaluación, que, como ya hemos dicho, a partir de ahora adoptaremos como fija. En el caso de la red simple, y debido a que no existe poliárbol de términos, en realidad sólo se utiliza la evaluación, ya que si la propagación se efectuara, la probabilidad a posteriori de cada término que no pertenece a la consulta coincidiría con la a priori, para todos los términos no instanciados.

Este conjunto de experimentos pretende estudiar el rendimiento combinado de los estimadores de distribuciones marginales con las funciones de probabilidad. Para ello se han probado todas las combinaciones de todos los estimadores de distribuciones marginales, menos *pp4* [5], con todas las funciones de probabilidad. El resto de los parámetros quedarán fijados a los siguientes valores: se empleará la diferencia de probabilidades (hecho que notaremos con un signo menos detrás del nombre de la función de probabilidad, por ejemplo, *fp1-*) y no se incorporarán los qf a las funciones de probabilidad (estos parámetros serán objeto de otros experimentos posteriores).

En cuanto al hecho de incorporar la diferencia de probabilidades del documento para efectuar la ordenación, hay que comentar que en la red simple esta acción equivale a tener en cuenta sólo los términos de cada documento que figuran en la consulta, ya que para los términos que no pertenezcan a la consulta las probabilidades a priori y a posteriori coinciden.

[5] La razón para excluir *pp4* es el hecho de que *pp2* y *pp4* ofrecen resultados muy parecidos. La influencia que ofrece el cociente $\frac{N-n_i}{N}$ en la expresión *pp4* es prácticamente nula, ya que en una colección con un número elevado de documentos, dicho cociente es aproximadamente 1 y, por tanto, las probabilidades devueltas por el estimador prácticamente son las que devuelve el estimador $pp2$.

Exp.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
pp1									
REL. REC.	91	26	55	82	53	50	47	52	54
M. 11PTS	0.4706	0.1144	0.2124	0.3098	0.2302	0.1964	0.2299	0.1977	0.1826
%C. 11PTS		-75.7	-54.9	-34.2	-51.1	-58.3	-51.2	-58.0	-61.2
pp2									
REL. REC.	91	25	84	82	84	86	82	90	89
M. 11PTS	0.4706	0.1130	0.3765	0.3098	0.4239	0.4034	0.3675	0.4297	**0.4516**
%C. 11PTS		-76.0	-20.0	-34.2	-9.9	-14.3	-21.9	-8.7	**-4.0**
pp3									
REL. REC.	91	25	84	82	80	85	84	84	92
M. 11PTS	0.4706	0.1131	0.3658	0.3098	0.4170	0.3935	0.3649	0.4236	0.4478
%C. 11PTS		-76.0	-22.3	-34.2	-11.4	-16.4	-22.5	-10.0	-4.8

Cuadro 3.5.: Batería I para la red simple de ADI: resultados para todas la funciones de probabilidad según los estimadores de las distribuciones marginales (*pp1, pp2 y pp3*), sin qf.

EXP.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
pp1									
REL. REC.	246	1	10	26	26	1	18	10	5
M. 11PTS	0.3768	0.0108	0.0291	0.0563	0.0657	0.0105	0.0495	0.0243	0.0168
%C. 11PTS		-97.1	-92.3	-85.1	-82.6	-97.2	-86.9	-93.6	-95.5
pp2									
REL. REC.	246	1	78	26	99	77	83	99	147
E. 11PTS	0.3768	0.0107	0.1685	0.0563	0.2186	0.1792	0.1954	0.2118	**0.3573**
%C. 11PTS		-97.2	-55.3	-85.1	-42.0	-52.4	-48.1	-43.8	**-5.2**
pp3									
REL. REC.	246	1	79	26	98	77	82	100	146
M. 11PTS	0.3768	0.0107	0.1660	0.0563	0.2180	0.1782	0.1935	0.2114	0.3570
%C. 11PTS		-97.2	-55.9	-85.1	-42.1	-52.7	-48.6	-43.9	-5.2

Cuadro 3.6.: Batería I para la red simple de CACM: resultados para todas la funciones de probabilidad según los estimadores de las distribuciones marginales (*pp1, pp2 y pp3*), sin qf.

Exp.	Smart	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
pp1									
Rel. Rec.	343	35	155	142	165	67	124	103	77
E. 11pts	0.2459	0.0476	0.1053	0.1141	0.1151	0.0614	0.0949	0.0861	0.0678
%C. 11pts		-80.7	-57.2	-53.6	-53.2	-75.0	-61.4	-65.0	-72.4
pp2									
Rel. Rec.	343	35	211	142	231	210	223	217	297
M. 11pts	0.2459	0.0467	0.1447	0.1141	0.1524	0.1319	0.1524	0.1501	**0.1997**
%C. 11pts		-81.0	-41.2	-53.6	-38.0	-46.3	-38.0	-39.0	**-18.8**
pp3									
Rel. Rec.	343	35	210	142	230	208	221	217	294
M. 11pts	0.2459	0.0473	0.1421	0.1141	0.1525	0.1314	0.1517	0.1492	0.1990
%C. 11pts		-80.7	-42.2	-53.6	-38.0	-46.6	-38.3	-39.3	-19.1

Cuadro 3.7.: Batería I para la red simple de CISI: resultados para todas la funciones de probabilidad según los estimadores de las distribuciones marginales (*pp1, pp2 y pp3*), sin qf.

Exp.	Smart	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
pp1									
Rel. Rec.	824	23	444	467	463	291	261	331	331
M. 11pts	0.4294	0.0106	0.2355	0.2325	0.2501	0.1241	0.1088	0.1706	0.1107
%C. 11pts		-97.5	-45.2	-45.9	-41.7	-71.1	-74.7	-60.3	-74.2
pp2									
Rel. Rec.	824	23	630	467	656	647	604	621	777
M. 11pts	0.4294	0.0105	0.3295	0.2325	0.3231	0.3381	0.2716	0.3219	**0.4070**
%C. 11pts		-97.5	-23.3	-45.9	-24.8	-21.3	-36.8	-25.0	**-5.2**
pp3									
Exp.	Smart	fp1	fp2	fp3	fp4	fp5	fp6	fp8	fp10
Rel. Rec.	824	23	625	467	655	634	596	620	776
M. 11pts	0.4294	0.0105	0.3251	0.2325	0.3205	0.3306	0.2683	0.3202	0.4045
%C. 11pts		-97.5	-24.3	-45.9	-25.4	-23.0	-37.5	-25.4	-5.8

Cuadro 3.8.: Batería I para la red simple de CRANFIELD: resultados para todas la funciones de probabilidad según los estimadores de las distribuciones marginales (*pp1, pp2 y pp3*), sin qf.

EXP.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
pp1									
REL. REC.	260	8	155	196	132	100	89	127	138
M. 11PTS	0.5446	0.0310	0.2774	0.3934	0.2486	0.1828	0.1691	0.2244	0.2342
%C. 11PTS		-94.3	-49.1	-27.8	-54.4	-66.4	-68.9	-58.8	-57.0
pp2									
REL. REC.	260	6	242	196	225	236	196	242	266
M. 11PTS	0.5446	0.0323	0.4848	0.3934	0.4361	0.4754	0.3687	0.4878	**0.5345**
%C. 11PTS		-94.1	-11.0	-27.8	-19.9	-12.7	-32.3	-10.4	**-1.8**
pp3									
REL. REC.	260	6	241	196	225	237	197	241	266
M. 11PTS	0.5446	0.0323	0.4840	0.3934	0.4354	0.4747	0.3666	0.4866	0.5319
%C. 11PTS		-94.1	-11.1	-27.8	-20.1	-12.8	-32.7	-10.7	-2.3

Cuadro 3.9.: Batería I para la red simple de MEDLARS: resultados para todas la funciones de probabilidad según los estimadores de las distribuciones marginales (*pp1, pp2 y pp3*), sin qf.

El comportamiento de la red simple en las cinco colecciones es sorprendentemente homogéneo. Se presenta de manera resumida en la tabla 3.5 para ADI, 3.6 para CACM, 3.7 para CISI, 3.8 para CRANFIELD y para MEDLARS en 3.9 (en el anexo B de esta memoria, figuran las tablas con los resultados completos, incluidas las curvas exhaustividad - precisión).

En general, esta conducta se puede calificar como mala, ya que en ningún caso se alcanza un porcentaje de cambio positivo con respecto a la media de los once puntos de exhaustividad de SMART.

De los valores mostrados en las tablas, podemos concluir que utilizando el estimador *pp2* los resultados mejoran sensiblemente con respecto a los otros dos; *pp1* es el peor con diferencia; *pp3* suele estar algo por encima de *pp1* y por debajo de *pp2*, igualándolo en algunos casos.

De manera generalizada, la función de probabilidad *fp10* rinde mucho más que sus compañeras, llegando a alcanzar porcentajes de cambio próximos a cero. Justo lo contrario que *fp1*, que posee la peor conducta recuperadora. A *fp10* le sigue normalmente, aunque a bastante distancia, la función *fp8*. El resto de funciones de probabilidad tiene una capacidad de recuperación muy similar, sin que haya ninguna que destaque considerablemente.

Por estas razones anteriores, para la siguiente batería de experimentos, y con objeto de reducir el espacio de búsqueda, se van a seleccionar únicamente *pp2* como estimador de distribuciones marginales y las funciones de probabilidad *fp8* y *fp10*.

3.6.2.2. Batería de experimentos II.

En la fase experimental anterior consideramos la diferencia de probabilidades y no se incluyeron la frecuencia de los términos en la consulta. Ahora, por tanto, estudiaremos el impacto en el rendimiento de la red simple cuando se modifique estos parámetros, e intentaremos dar respuesta a las dos cuestiones siguientes.

EXP.	SMART	0, –	0	1, –	1
fp8					
REL. REC.	91	90	88	91	90
M. 11PTS	0.4706	0.4297	0.4508	0.4356	0.4552
%C. 11PTS		-8.7	-4.2	-7.4	-3.3
fp10					
REL. REC.	91	89	90	90	91
M. 11PTS	0.4706	0.4516	0.4707	0.4575	**0.4709**
%C. 11PTS		-4.0	0.0	-2.8	**0.1**

Cuadro 3.10.: Batería II para la red simple de ADI: *pp2* y *fp10, fp8*, con y sin *qf*, con y sin diferencia de probabilidades.

EXP.	SMART	0, –	0	1, –	1
fp8					
REL. REC.	246	99	100	100	102
M. 11PTS	0.3768	0.2118	0.2190	0.2179	0.2241
%C. 11PTS		-43.8	-41.9	-42.2	-40.5
fp10					
REL. REC.	246	147	142	146	143
M. 11PTS	0.3768	0.3573	**0.3582**	0.3425	0.3435
%C. 11PTS		-5.2	**-4.9**	-9.1	-8.8

Cuadro 3.11.: Batería II para la red simple de CACM: *pp2* y *fp8, fp10*, con y sin *qf*, con y sin diferencia de probabilidades.

- ¿Es preferible la incorporación de los pesos de los términos que aparecen en la consulta (en las tablas notado como 1) a la función de probabilidad, o por el contrario es mejor no utilizarlos (notado como 0)?
- ¿Se alcanza un mayor rendimiento utilizando la diferencia de probabilidades del documento ($-$) o sólo la probabilidad a posteriori[6]?

Además, se desea confirmar el comportamiento de las funciones de probabilidad seleccionadas en la fase anterior: *fp8* y *fp10*.

A la luz del contenido resumido de las tablas 3.10, 3.11, 3.12, 3.13 y 3.14, podemos afirmar que:

- En la red simple, es mejor no utilizar la diferencia de probabilidades del documento y establecer la ordenación sólo mediante la probabilidad a posteriori. En todas las colecciones se cumplen que cuando intervienen todos los términos del documento en la sumatoria de la función de probabilidad, se alcanza un mayor rendimiento que en el caso en el que sólo intervienen los que están en la consulta.

[6]Cuando el identificador de la función de probabilidad correspondiente no vaya seguido del signo $-$ indicaremos que sólo se ha utilizado la probabilidad a posteriori y no la diferencia.

Exp.	Smart	0, –	0	1, –	1
fp8					
Rel. Rec.	343	217	257	264	300
M. 11pts	0.2459	0.1501	0.1823	0.1892	0.2160
%C. 11pts		-39.0	-25.9	-23.1	-12.2
fp10					
Rel. Rec.	343	297	312	349	369
M. 11pts	0.2459	0.1997	0.2206	0.2520	**0.2642**
%C. 11pts		-18.8	-10.3	2.5	**7.4**

Cuadro 3.12.: Batería II para la red simple de CISI: *pp2* y *fp8, fp10*, con y sin *qf*, con y sin diferencia de probabilidades.

Exp.	Smart	0, –	0	1, –	1
fp8					
Rel. Rec.	824	621	708	624	709
M. 11pts	0.4294	0.3219	0.3706	0.3223	0.3695
%C. 11pts		-25.0	-13.7	-24.9	-13.9
fp10					
Rel. Rec.	824	777	816	779	825
M. 11pts	0.4294	0.4070	**0.4323**	0.4069	0.4309
%C. 11pts		-5.2	**0.7**	-5.3	0.4

Cuadro 3.13.: Batería II para la red simple de CRANFIELD: *pp2* y *fp8, fp10*, con y sin *qf*, con y sin diferencias de probabilidades.

Exp.	Smart	0, –	0	1, –	1
fp1					
Rel. Rec.	260	242	253	235	246
M. 11pts	0.5446	0.4878	0.5262	0.4895	0.5193
%C. 11pts		-10.4	-3.4	-10.1	-4.6
fp10					
Rel. Rec.	260	266	269	258	264
M. 11pts	0.5446	0.5345	**0.5552**	0.5272	0.5458
%C. 11pts		-1.8	**1.9**	-3.2	0.2

Cuadro 3.14.: Batería II para la red simple de MEDLARS: *pp2* y *fp8, fp10*, con y sin *qf*, con y sin diferencias de probabilidades.

- *fp10* es la función de probabilidad que mejor rendimiento tiene.
- Las colecciones ADI y CISI tienen mejor rendimiento cuando se emplean los *qf*. Al contrario ocurre con el resto, donde es mejor no incluirlos en la evaluación del documento. En cualquier caso, únicamente para CISI, y en menor medida para CACM, las diferencias entre un método y otro son importantes.

En esta fase de la experimentación, cuatro de las cinco colecciones consiguen un porcentaje de cambio superior a SMART. Sólo CACM sigue en valores negativos.

A partir de este momento, tomaremos *fp10* como la función de probabilidad con la que seguir realizando pruebas y *pp2* el estimador de probabilidades marginales.

3.6.3. Experimentación con la red aumentada.

Pasemos seguidamente a determinar el rendimiento de la red aumentada, para lo cual, y como hicimos con la red simple, expondremos el conjunto de parámetros en los que estamos interesados:

1. *Estimación de distribuciones de probabilidad:*
   - Estimadores de distribuciones de probabilidad marginales: únicamente consideramos *pp2*.
   - Estimadores de distribuciones de probabilidad condicionada: compararemos entre *pc-mv* y *pc-J*[7].
2. *Propagación:*
   - Funciones de probabilidad: *fp10*.
   - Uso de *qf* (1) en ADI y en CISI. En el resto, no se usa el *qf* (0).
   - Instanciación total o parcial de evidencias.
   - Ordenación de documentos según su probabilidad a posteriori o la diferencia de la a posteriori y la a priori (-).

3.6.3.1. Batería de experimentos I: determinación del mejor estimador de distribuciones condicionadas y confirmación del uso de los *qf* de los términos de la consulta.

La primera fase de prueba con la red aumentada centra su interés en determinar cuál de los dos estimadores de las distribuciones de probabilidad condicionada en los nodos término, el de máxima verosimilitud *pc-mv* o Jaccard *pc-J*, es el idóneo para este modelo. De manera paralela,

Exp.	SMART	*pc-mv*, 0	*pc-mv*, 1	*pc-J*, 0	*pc-J*, 1
ADI					
REL. REC.	91	90	91	92	93
M. 11PTS	0.4706	0.3501	0.4509	0.4130	**0.4613**
%C. 11PTS		-25.6	-4.2	-12.2	**-2.0**
CACM					
REL. REC.	246	223	247	230	244
M. 11PTS	0.3768	0.3582	0.4041	0.3759	**0.4046**
%C. 11PTS		-4.9	7.2	-0.2	**7.3**
CISI					
REL. REC.	343	277	318	282	318
M. 11PTS	0.2459	0.1948	0.2290	0.2007	**0.2301**
%C. 11PTS		-20.8	-6.9	-18.4	**-6.4**
CRANFIELD					
REL. REC.	824	812	814	826	809
M. 11PTS	0.4294	0.4220	0.4136	**0.4314**	0.4116
%C. 11PTS		-1.7	-3.7	**0.5**	-4.1
MEDLARS					
REL. REC.	260	288	265	292	266
M. 11PTS	0.5446	0.6134	0.5836	**0.6200**	0.5792
%C. 11PTS		12.6	7.2	**13.8**	6.4

Cuadro 3.15.: Batería I para la red aumentada de las cinco colecciones: *pp2, fp10, pc-mv vs pc-J, con y sin qf*.

comprobaremos si se mantiene la misma línea general que se dio en la red simple con respecto al uso o no de los *qf* de los términos de la consulta.

Como conclusión al estudio de los resultados resumidos expuestos en la tabla 3.15, podemos indicar que el estimador de distribuciones de probabilidad condicionadas de Jaccard consigue, a lo largo de todas las colecciones, unos porcentajes de cambio mayores que su análogo de máxima verosimilitud. Por otro lado, el comportamiento con respecto al uso o no de la técnica de la replicación (utilización o no de los *qf*) se mantiene casi igual que con la red simple: con ADI y CISI es mejor utilizarla; con CRANFIELD y MEDLARS, no; y con CACM el porcentaje de cambio a favor de la replicación supera al conseguido cuando no se utiliza (esta conducta se mantiene en posteriores experimentos), dándole la vuelta al comportamiento mostrado en pruebas pasadas.

Tanto para CACM como MEDLARS, la red aumentada con los valores *pp2, pc-J, fp10* y con replicación para la primera colección, y sin replicación para la segunda, establecen nuevos máximos en cuanto al porcentaje de cambio obtenido. La eficacia recuperadora en estas dos colecciones se ve incrementada claramente cuando se incorporan las relaciones entre términos en la subred de términos. Justo al contrario ocurre con ADI y CISI, que pierden en calidad recuperadora con respecto a la red simple. Por otro lado, podríamos decir que CRANFIELD mantiene la misma tónica en ambas redes.

[7] Para ver el rendimiento detallado del estimador bayesiano *pc-eb*, refiérase al trabajo [CFH00]

EXP.	SMART	fp10	fp10-
ADI (1)			
REL. REC.	91	93	91
M. 11PTS	0.4706	**0.4613**	0.4581
%C. 11PTS		**-2.0**	-2.7
CACM (1)			
REL. REC.	246	244	242
M. 11PTS	0.3768	**0.4046**	0.3996
%C. 11PTS		**7.3**	6.0
CISI (1)			
REL. REC.	343	318	309
M. 11PTS	0.2459	**0.2301**	0.2299
%C. 11PTS		**-6.4**	-6.5
CRANFIELD (0)			
REL. REC.	824	826	847
M. 11PTS	0.4294	0.4314	**0.4421**
%C. 11PTS		0.5	**2.9**
MEDLARS (0)			
REL. REC.	260	292	293
M. 11PTS	0.5446	0.6200	**0.6407**
%C. 11PTS		13.8	**17.6**

Cuadro 3.16.: Batería II para la red aumentada de todas las colecciones: *pp2, pc-J, qf* según la colección y *fp10 vs fpl0-*.

Atendiendo a estos resultados, optaremos por continuar usando siempre *pc-J*, e inclusión de los *qf* de términos de la consulta sólo en ADI, CACM y CISI. En el resto de colecciones no se usará esta técnica.

3.6.3.2. Batería de experimentos II: ordenación de documentos mediante probabilidad a posteriori o mediante la diferencia de probabilidades.

Este segundo experimento va a determinar qué es mejor: ordenar los documentos según únicamente su probabilidad a posteriori o, por el contrario, hacer la ordenación basándose en la diferencia existente entre la a posteriori y la a priori del documento. Para comprobarlo, se ha efectuado una recuperación ordenando por la diferencia, cuyos resultados abreviados se presentan en la tabla 3.16. Para cada colección, indicamos junto al nombre de la misma si se considera el *qf* (1) o no se considera (0).

CRANFIELD y MEDLARS tienen un mejor rendimiento cuando se considera la diferencia de probabilidades. De hecho, ambas alcanzan el mejor porcentaje de cambio hasta ahora. Para las tres primeras, ADI, CACM y CISI, su comportamiento es ligeramente peor al usar la diferencia de probabilidades, aunque en posteriores experimentos se demuestra que esta conducta es puntual, mejorando sensiblemente cuando se ordenan los documentos por la diferencia. Así, elegiremos para la tercera y última fase la función de probabilidad *fp10-*.

EXP.	SMART	$\alpha = 0{,}2$	$\alpha = 0{,}5$	$\alpha = 0{,}8$
ADI (1)				
REL. REC.	91	90	91	91
M. 11PTS	0.4706	**0.4628**	0.4598	0.4607
%C. 11PTS		**-1.7**	-2.3	-2.2
CACM (1)				
REL. REC.	246	220	228	236
M. 11PTS	0.3768	0.3838	0.3914	**0.4005**
%C. 11PTS		1.8	3.8	**6.2**
CISI (1)				
REL. REC.	343	308	287	268
M. 11PTS	0.2459	**0.2291**	0.2050	0.1917
%C. 11PTS		**-6.9**	-16.7	-22.1
CRANFIELD (0)				
REL. REC.	824	819	839	842
M. 11PTS	0.4294	0.4248	0.4344	**0.4391**
%C. 11PTS		-1.1	1.1	**2.2**
MEDLARS (0)				
REL. REC.	260	287	287	293
M. 11PTS	0.5446	0.6428	**0.6443**	0.6426
%C. 11PTS		18.0	**18.3**	17.9

Cuadro 3.17.: Batería III para la red aumentada de todas las colecciones: *pp2, pc-J, fp10-* e instanciación parcial.

3.6.3.3. Batería de experimentos III: Instanciación total o uso de evidencias parciales.

El último parámetro que queda por probar es el uso de evidencias parciales. En este caso, la ecuación (3.12)

$$p(t_i \mid T_i \in Q) = \alpha + (1-\alpha)\frac{tf_{iQ} \cdot idf_i}{max_{T_j \in Q} tf_{jQ} \cdot idf_j}$$

nos determina el valor de la probabilidad a posteriori que se desea conseguir, el cual depende de una combinación convexa de la calidad del término, siendo α el parámetro que la controla. Por tanto, dependiendo de lo que valga α, la probabilidad serás más cercana a uno o tenderá más al peso del término.

En esta fase comprobaremos si es útil la puesta en práctica de la instanciación parcial, y al mismo tiempo, cuál es el mejor valor para α. Para ello se realizarán tres experimentos, uno con cada uno de los valores 0,2, 0,5 y 0,8, con objeto de cubrir el abanico de valores bajos, medios y altos.

La tabla 3.17 muestra los resultados de esta fase experimental con la red aumentada. De nuevo MEDLARS alcanza un nuevo máximo en el porcentaje de cambio con respecto a SMART. De la observación de esta tabla se llegan a distinguir claramente tres conductas:

1. Los porcentajes de cambio mejoran conforme aumenta el valor de α. Esta situación se da

en CACM y CRANFIELD.

2. Los porcentajes de cambio mejoran conforme disminuye el valor de α, comportamiento que ofrecen ADI y CISI.

3. No se da ninguna de estas dos circunstancias: MEDLARS.

En la primera situación, los resultados obtenidos por el valor 0,8 de α están algo por debajo de los conseguidos con instanciación total. Parece, por tanto, que lo ideal es instanciar los términos de la consulta a relevante. Sin embargo, en la segunda, el porcentaje de cambio con $\alpha = 0{,}2$ es ligeramente mejor en ADI que el de la instanciación total y aproximadamente igual que en CISI.

Resumiendo, parece que para la red aumentada es conveniente utilizar, independientemente del tipo de colección, el estimador de distribuciones de probabilidad condicionadas de Jaccard (*pc-J*) y la función de probabilidad *fp10* con las diferencias entre probabilidades a posteriori y a priori de cada documento. Para ADI, CACM y CISI es adecuado usar el *qf* en los nodos de la consulta, cosa que no ocurre con CRANFIELD y MEDLARS. Por otro lado, y excepto para MEDLARS, la instanciación total alcanza mejores porcentajes de cambio. Hasta ahora, ADI y CISI alcanzan un mejor rendimiento con la red simple y las tres colecciones restantes lo hacen con la red aumentada.

3.6.4. Experimentación con la red mixta.

Finalizaremos la exposición de los resultados conseguidos en la fase experimental con el último modelo desarrollado: la red mixta. El conjunto de parámetros, y sus valores (muchos de ellos se heredan de las etapas anteriores) con los que podemos efectuar las pruebas son los siguientes:

1. *Estimación de distribuciones de probabilidad:*
   - Estimadores de distribuciones de probabilidad marginales: *pp2*.
   - Estimadores de distribuciones de probabilidad condicionada: *pc-J*.

2. *Propagación:*
   - Funciones de probabilidad: *fp10, fp10c, fp10d y fp10e*.
   - Inclusión del *qf* de los términos de la consulta (1) en ADI, CACM y CISI y sin utilizarlos (0) en CRANFIELD y MEDLARS.
   - Instanciación total o parcial de evidencias.
   - Ordenación de documentos según su probabilidad a posteriori o la diferencia de la a posteriori y la a priori (-).

Por tanto, nos planteamos determinar la calidad recuperadora del S.R.I. considerando estos distintos parámetros, para lo cual se han diseñado las baterías que describimos a continuación.

EXP.	SMART	fp10-	fp10	fp10c-	fp10c	fp10d-	fp10d	fp10e-	fp10e
ADI (1)									
REL. REC.	91	88	88	73	68	84	82	88	88
M. 11PTS	0.4706	0.4470	0.4459	0.3907	0.3813	**0.4501**	0.4477	0.4470	0.4464
%C. 11PTS		-5.1	-5.3	-17.0	-19.0	**-4.4**	-4.9	-5.1	-5.2
CACM (1)									
REL. REC.	246	240	241	197	199	219	224	240	241
M. 11PTS	0.3768	0.3956	0.3972	0.2488	0.2517	0.3667	0.3708	0.3956	**0.3988**
%C. 11PTS		4.9	5.4	-34.0	-33.3	-2.7	-1.6	4.9	**5.8**
CISI (1)									
REL. REC.	343	340	344	293	299	332	336	340	344
M. 11PTS	0.2459	0.2521	0.2506	0.2107	0.2111	**0.2529**	0.2528	0.2521	0.2506
%C. 11PTS		2.5	1.9	-14.4	-14.2	**2.8**	2.8	2.5	1.9
CRANFIELD (0)									
REL. REC.	824	831	813	704	717	836	829	829	809
M. 11PTS	0.4294	0.4347	0.4257	0.3556	0.3539	**0.4434**	0.4389	0.4342	0.4258
%C. 11PTS		1.2	-.9	-17.2	-17.6	**3.2**	2.2	1.1	-.9
MEDLARS (0)									
REL. REC.	260	285	277	288	288	287	287	285	277
M. 11PTS	0.5446	0.6108	0.5911	0.6141	0.6075	**0.6281**	0.6188	0.6108	0.5910
%C. 11PTS		12.1	8.5	12.7	11.5	**15.3**	13.6	12.1	8.5

Cuadro 3.18.: Batería I para la red mixta de todas las colecciones: *pp2, pc-J*, uso o no del *qf* dependiendo de la colección. Comparación de las funciones de probabilidad *fp10, fp10c, fp10d, fp10e* y sus variaciones con las diferencias.

3.6.4.1. Batería de experimentos I: determinación de la mejor función de probabilidad.

La primera batería experimental diseñada para la red mixta tiene como objetivo evaluar la efectividad general del modelo y determinar la función de probabilidad que alcanza un mayor rendimiento a lo largo de las cinco colecciones. Para ello, se han efectuado recuperaciones con *fp10, fp10c, fp10d* y *fp10e*, además de las análogas con las diferencias de probabilidades para realizar la ordenación.

A la luz de los resultados de la tabla 3.18, en la mayoría de los casos la versión que contempla la diferencia de probabilidades alcanza una mayor efectividad que su versión sencilla correspondiente. Sistemáticamente, las funciones *fp10c* y *fp10c-* son las peores (recuérdese que con estas funciones de probabilidad se simula a una red en la que no se incluyen los términos aislados). Esto implica que la inclusión en el modelo de los nodos aislados es importante. En todas las colecciones, salvo en CACM, donde domina *fp10e*, la función *fp10d-* es la que llega a conseguir un mejor resultado (aunque en CISI la diferencia con su homóloga *fp10d* es ínfima).

En este caso, la inclusión de los nodos aislados de manera dinámica, es decir, dependiendo de la consulta, aporta incrementos en el rendimiento del modelo. Por tanto, como conclusión a esta fase, tomaremos para la siguiente etapa experimental la función de probabilidad *fp10d-*.

De nuevo, como conclusión general, podemos decir que, a la luz de los resultados ofrecidos en la tabla 3.18, la capacidad recuperadora de la red mixta depende de la colección con que

	SMART	$\alpha = 0{,}2$	$\alpha = 0{,}5$	$\alpha = 0{,}8$
ADI (1)				
REL. REC.	91	82	83	83
M. 11PTS	0.4706	0.4428	**0.4503**	0.4474
%C. 11PTS		-6.0	**-4.4**	-5.0
CACM (1)				
REL. REC.	246	220	232	238
M. 11PTS	0.3768	0.3817	0.3904	**0.3984**
%C. 11PTS		1.3	3.6	**5.7**
CISI (1)				
REL. REC.	343	339	333	308
M. 11PTS	**0.2459**	0.2551	0.2537	0.2381
%C. 11PTS		**3.7**	3.1	-3.2
CRANFIELD (0)				
REL. REC.	824	803	833	840
M. 11PTS	0.4294	0.4144	0.4290	**0.4387**
%C. 11PTS		-3.5	-.1	**2.1**
MEDLARS (0)				
REL. REC.	260	287	289	286
M. 11PTS	0.5446	0.6246	**0.6298**	0.6272
%C. 11PTS		14.6	**15.6**	15.1

Cuadro 3.19.: Batería II para la red mixta de todas las colecciones: *pp2, pc-J, fp10d- (salvo CACM, fp10e) e instanciación parcial*.

trabajemos. Así, para ADI, se consiguen valores que están por debajo de los ofrecidos por las redes simple y aumentada. Los generados por CACM y MEDLARS superan a la red simple pero están por debajo de los de la red aumentada, perdiendo el S.R.I. aproximadamente un 2 % de capacidad recuperadora. Para CISI, es justo al contrario: mejora la aumentada, pero siguen siendo mejores los resultados obtenidos con la red simple. Por último, es la red mixta de CRANFIELD la que consigue mejoras sobre los otros dos modelos, estableciendo un nuevo mejor porcentaje de cambio.

3.6.4.2. Batería de experimentos II: Instanciación total o uso de evidencias parciales.

Para concluir, comprobaremos la sensibilidad de la red mixta con respecto al uso de las evidencias parciales. Al igual que hicimos con la red aumentada, se realizarán pruebas con $\alpha = 0{,}2, 0{,}5$ y $0{,}8$.

Una de las primeras conclusiones que se sacan haciendo el análisis de los resultados incluidos en la tabla 3.19 es que el comportamiento por colecciones conforme aumenta el valor de α es muy parecido al que se consiguió con la red aumentada. Así, CACM y CRANFIELD aumentan su rendimiento conforme se incremente α. CISI, por el contrario, rinde más cuando este parámetro disminuye. Con la red aumentada, también ocurría con ADI, aunque ahora su conducta es un tanto irregular. Igual que sucede con MEDLARS.

Para CACM y CRANFIELD, el porcentaje de cambio cuando $\alpha = 0{,}8$ es el mejor obtenido en la experimentación con la red mixta. Igual ocurre con CISI con $\alpha = 0{,}2$. En las otras dos colecciones, ADI y MEDLARS, el rendimiento máximo se alcanzan con $\alpha = 0{,}5$.

Otra conclusión a la que podemos llegar es que las colecciones ADI, CRANFIELD y MEDLARS, con la red mixta y evidencias parciales, no rinden a la altura de la red aumentada y evidencias parciales, estando algo por debajo en los porcentajes de cambio. CACM también está por debajo, pero muy ligeramente. CISI es la única que mejora los resultados conseguidos por la red aumentada.

En cualquier caso, el comportamiento de la red mixta en todos los experimentos realizados es razonable, lo que nos hará pensar que la selección de términos es de utilidad, sobre todo cuando se trabaje con grandes colecciones.

3.6.5. Conclusiones de la experimentación.

En la tabla 3.20 presentamos las distintas combinaciones de parámetros que hacen que la capacidad recuperadora del modelo alcance su máximo valor para cada colección. Y en el B se muestran las gráficas Exhaustividad - Precisión correspondientes.

COLECCIÓN	TIPO DE RED	PARÁMETROS	%CAMBIO
ADI	Red simple	*pp2, con qf, fp10*	0.1
CACM	Red aumentada	*pp2, pc-J, con qf, fp10*	7.3
CISI	Red simple	*pp2, con qf, fp10*	7,4
CRANFIELD	Red mixta	*pp2, pc-J, sin qf, fp10d-*	3.2
MEDLARS	Red aumentada	*pp2, pc-J, sin qf, fp10-, instanciación parcial* ($\alpha = 0{,}5$)	18.3

Cuadro 3.20.: Mejores resultados obtenidos en cada colección.

Como conclusiones finales a esta experimentación podemos destacar las siguientes:

- Las tres topologías para la subred de términos se presentan como un amplio abanico donde poder elegir aquélla que más se adecue a la colección con la que se esté trabajando. No hay ninguna topología que sistemáticamente ofrezca un mejor rendimiento que las otras. Además, es difícil conseguir un modelo que se comporte de forma estable con las distintas colecciones estudiadas. Esto es un hecho, por otra parte, usual cuando se estudian los valores obtenidos por otros S.R.I.

- No es aconsejable el uso de algoritmos aproximados basados en técnicas de simulación para propagar en las redes bayesianas documentales. La experimentación nos muestra que los resultados obtenidos no se acercan a los valores proporcionados a la propagación exacta.

- Si exceptuamos a la colección ADI, los resultados obtenidos muestran la viabilidad de un S.R.I. basado en redes bayesianas.

- El estimador de las distribuciones de probabilidad marginal *pp2* se ha mostrado como el más idóneo, independientemente de la colección, indicando así que es preferible utilizar una probabilidad a priori (y pequeña para el valor relevante) igual para todos los nodos sin padres.

- El estimador de distribuciones de probabilidad condicionada con la que el S.R.I. alcanza una mayor capacidad recuperadora, independientemente de la colección, es *pc-J*.

- En la red simple, la función de probabilidad que destaca, en cuanto a rendimiento, con respecto a las demás diseñadas, es la *fp10*.

- La diferencia de las probabilidades a posteriori y a priori del documento para generar la ordenación de documentos es una técnica cuyo uso parece adecuado para todas las colecciones.

- Puede ser interesante el modificar la importancia de los pesos asociados a los términos en la red bayesiana documental. Así, tanto el uso de los *qf* de los términos del documento que aparecen en la consulta, como el considerar evidencias parciales, se ha demostrado beneficioso para algunas colecciones. Además, el comportamiento parece ser independiente de la topología utilizada.

3.7. Comparativa con otros modelos de recuperación basados en redes bayesianas.

Para concluir este capítulo vamos a exponer las principales diferencias y similitudes de los modelos de recuperación que se han desarrollado con los tres principales existentes en la literatura: *Inference Network, Belief Network* y el de Ghazfan y col.

- Todos los modelos poseen aproximadamente el mismo tipo de nodos. La única diferencia más significativa es que nuestros modelos no poseen nodo consulta y, por tanto, la red es estática con respecto a ésta. La llegada de una nueva consulta al S.R.I. se trata instanciando los términos que contiene a relevantes en lugar de crear una subred de consultas para cada una que se someta al sistema.

- En cuanto a la orientación de los arcos, en nuestro modelo, al igual que en el Belief network y en el de Ghazfan y col., se hace de los nodos término a los nodos documento. Por el contrario, en el Inference Network se pone en práctica justo de manera contraria. Croft y Turtle lo justifican desde el punto de vista de la causalidad. El resto lo hacemos por razones puramente operacionales relacionadas con la propagación y la semántica del proceso de propagación, ya que resulta más clara de esta forma.

- El problema del gran número de padres que poseen los nodos documento lo resolvemos en el modelo de red bayesiana documental utilizando las funciones de probabilidad. Sólo Gazfhan y col. plantean también este problema y lo solventan agrupando los términos de cada documento en diferentes grupos. A cada grupo se le asocia un nodo y se enlaza con el nodo documento correspondiente y con los nodos términos que contienen. De esta manera actúan de capa intermedia y reducen el tamaño de las distribuciones de probabilidad.

- Tanto la red aumentada como la mixta son los dos únicos modelos que incorporan relaciones directas entre términos, aumentando así la expresividad de la red. El resto de modelos no consideran esta característica.

- La propagación se hace una única vez, como también ocurre en los modelos Belief Network y el de Ghazfan y col., instanciando la consulta y propagando hacia los documentos, calculando así la probabilidad de relevancia de cada documento dada la consulta. Croft y Turtle instancian un documento sucesivamente y propagan hacia el nodo necesidad de información, calculando así la probabilidad de que la necesidad de información quede satisfecha dado un nodo. Esta forma de proceder requiere tantas propagaciones como documentos existan en la colección.

- Los modelos Inference network y Belief network han sido diseñados para simular el comportamiento de los modelos booleano y probabilístico, además del vectorial, en el caso del segundo. El modelo de la red bayesiana documental no simula a ningún otro aunque su extensión no plantea dificultades.

Modelo	Colección	M. 11PTS	M. 11PTS - R.B.D.	%C
Inference Network	ADI	0.2154	0.4709	113.62
Inference Network	CACM	0.3740	0.4046	8.18
Ghazfan y col.	ADI	0.5033	0.4709	-6.44
Ghazfan y col.	CACM	0.3730	0.4046	8.47

Cuadro 3.21.: Comparativa de resultados con otros modelos basados en redes bayesianas.

- En los modelos Inference Network y Belief Network, la propagación se hace evaluando funciones que devuelven la probabilidad a posteriori. Ghazfan y col. aplican un algoritmo de propagación exacto (modifican el del paso de mensajes de Pearl para evitar que los mensajes discurran de manera infinita por los diferentes ciclos del grafo). En nuestro modelo, se aplica la propagación exacta del algoritmo de paso de mensajes en la subred de términos y se extiende a la subred de documentos mediante la evaluación de las funciones de probabilidad en los documentos.

- En cuanto a rendimiento, las comparaciones sólo se pueden establecer con el modelo Inference Network y con el de Ghazfan y col. ya que presentan resultados con las colecciones ADI y CACM. Con respecto al de Ribeiro y Reis, al trabajar con una colección distinta a las que nosotros hemos experimentado, no es posible llevar a cabo dicha comparativa.

  La tabla 3.21 muestra las medias de la precisión para los once puntos de exhaustividad (M. 11PTS), tanto para CACM y ADI, de los dos modelos. Estas medidas se han obtenido de [Tur90, GIS96]. Además, presentamos la misma media (M. 11PTS - R.B.D.) para el mejor valor conseguido con el modelo de red bayesiana documental con esas dos colecciones y el porcentaje de cambio que representa nuestra medida con respecto a la suya.

  Como se puede apreciar, el rendimiento depende claramente de la colección que se esté utilizando. Nuestro modelo de red bayesiana documental es superior en ADI cuando se compara con el modelo Inference Network e inferior con esa misma colección analizando los resultados del de Ghazfan y col. Con respecto a la CACM, se supera a ambos. Estas comparaciones hay que hacerlas con reserva, pues las condiciones experimentales pueden ser diferentes en cada modelo.

4. Realimentación de relevancia para los modelos de R.I. basados en redes bayesianas.

En este capítulo vamos a presentar los métodos de realimentación de relevancia desarrollados para los modelos basados en las diferentes redes bayesianas documentales. Previamente a esta exposición, se hará una breve introducción a esta técnica de modificación de la consulta, mostrando cómo se lleva a cabo en el conocido modelo del Espacio Vectorial. Seguidamente, comentaremos la manera en que los dos principales modelos de recuperación de información basados en redes Bayesianas ponen en práctica la realimentación.

El método que en este capítulo introducimos se presenta como una alternativa clara y novedosa a los ya existentes, ya que está apoyado en el concepto de evidencias parciales, concepto nunca antes utilizado en este campo. Ofrece, además, niveles de rendimiento que son totalmente comparables a los que se alcanzan con los modelos de redes alternativos, así como con los conseguidos por los clásicos del Espacio Vectorial y del Probabilístico.

4.1. Introducción a la realimentación de relevancia.

Clasificada como un método de modificación o reformulación de la consulta [SM83], *la realimentación de relevancia* es una técnica interactiva e iterativa para construir una consulta de mayor calidad que la original. El objetivo principal que se persigue, por tanto, es ayudar al usuario a obtener una representación más exacta de su necesidad de información real, con lo cual se recuperarán más documentos relevantes que satisfagan a éste.

El procedimiento en que está basada la realimentación es sencillo: el usuario inspecciona los documentos que el S.R.I. le ha devuelto tras realizar la recuperación para una consulta inicial, y decide cuáles son relevantes o no para esa consulta. Esta información, denominada *juicios de relevancia*, será usada por el S.R.I. para construir una nueva consulta, que será sometida posteriormente a recuperación. El carácter iterativo viene dado porque este proceso se puede reiterar tantas veces como lo considere oportuno el interesado, hasta que sus necesidades queden totalmente satisfechas. El aspecto interactivo queda determinado por el diálogo continuo entre el usuario y el S.R.I.

Tal y como Salton y Buckley señalan en [SB90], hay tres razones fundamentales por las que la realimentación se presenta como una técnica muy atractiva para el usuario:

- Éste no tiene conocimiento de cómo el S.R.I. construye la nueva consulta. Sólo se limita a ofrecer sus juicios de relevancia.
- La búsqueda global de todos los documentos relevantes se sustituye por una secuencia de varias búsquedas más simples que se acercan gradualmente al fin original.
- El usuario posee una herramienta poderosa con la que puede dar mucha importancia a los términos que aparecen en los documentos relevantes y quitársela a aquéllos que aparecen en los no relevantes.

En la mayoría de los S.R.I. que tengan la opción de efectuar la realimentación, son los términos que pertenecen a los documentos que han sido clasificados como relevantes o no relevantes los que realmente se usan para ponerla en práctica, en lugar de los propios documentos. Así, y teniendo en cuenta esta característica, está compuesta realmente por dos técnicas diferentes, pero complementarias a su vez [Har92b]:

1. *Repesado de la consulta original*: la importancia de los términos que figuran en la consulta inicial se modifica de acuerdo con su aparición en documentos relevantes o no. De esta manera, se incrementará la posición en la ordenación de documentos por relevancia de aquéllos que no han sido recuperados y que contienen los términos repesados positivamente y se disminuirá la situación de los que contengan los términos que pierdan peso.
2. *Expansión de la consulta*: la consulta original se ve aumentada por la inclusión de otros términos que no figuran en la cuestión primitiva, pero que sí lo hacen en documentos declarados por el usuario como relevantes o no. Así, se permite que otros documentos que no tienen intersección con la consulta original, pero que pueden resultar interesantes para el usuario, asciendan o desciendan en la ordenación por relevancia (en este último caso, si claramente contienen términos que no son útiles).

Spink y Losee han publicado un detallado estudio sobre la realimentación de relevancia [SL96] en el que tratan, entre otras cuestiones, cómo ponen en práctica en los principales modelos de recuperación esta técnica de modificación de consultas. Nosotros, con objeto de profundizar un poco más en ella y centrar así la exposición posterior de nuestro método, seguidamente trataremos la realimentación de relevancia en el modelo del Espacio Vectorial.

Como sabemos, este modelo representa los documentos y las consultas como vectores que contienen los pesos de los términos por los que han sido indexados (si no aparece algún término, el peso asignado es 0). Así, la realimentación se efectúa mediante la mezcla de los vectores de los documentos con el de la consulta, dando lugar, una vez terminado el proceso, a un nuevo vector que hará las veces de nueva consulta para el S.R.I.

Veamos seguidamente cómo funcionan las dos técnicas que componen la realimentación de relevancia en el modelo del Espacio Vectorial:

- *Repesado de la consulta original*: si un término aparece en un documento calificado como relevante, y lo hace también en el vector de la consulta, entonces sus correspondientes pesos se suman y se asignan a la posición que le corresponde a ese término en un nuevo vector. Con ello se premia a los términos de la consulta que aparecen en documentos relevantes, reforzando así su peso. En caso de que dicho término se dé en un documento no relevante, su peso será restado al del mismo término en la consulta. En este caso se produce una penalización del término de la consulta, quitándole peso, o lo que es lo mismo, dándole menos importancia.

- *Expansión de la consulta*: si un término aparece en un documento relevante y no lo hace en la consulta original, entonces su peso en el documento se asigna directamente al peso del término en el nuevo vector. De esta manera, la consulta se ve reforzada por términos potencialmente interesantes. Por otro lado, si el término ha sido visto en un documento no relevante, se añade a la nueva consulta pero con peso negativo. Ésta es la forma de indicar que el término no es bueno para alcanzar los objetivos del usuario. Si un término perteneciente a un documento, relevante o no ya ha sido observado previamente, y por tanto, ya ha sido añadido a la nueva consulta, se procederá a sumar o restar sus pesos, según sea el caso.

La siguiente fórmula representa numéricamente, y en terminología del Espacio Vectorial, la descripción anterior [Har92]:

$$Q_1 = Q_0 + \sum_{j=1}^{|R|} D_j - \sum_{j=1}^{|\bar{R}|} D_j \tag{4.1}$$

donde, R y $\bar{R}$ son los conjuntos de documentos relevantes y no relevantes conocidos, respectivamente; $\mid R \mid$ y $\mid \bar{R} \mid$ sus cardinales; D_j el vector de un documento cualquiera de R o $\bar{R}$; y por último, Q_0 y Q_1 los vectores de la consulta original y la nueva consulta generada, por este orden. Este método se denomina *Ide Regular*.

Una variante de este método anterior es la conocida como *Ide Dec-Hi*, la cual sólo tiene en cuenta el documento no relevante que se sitúe en la posición más alta en la ordenación de documentos recuperados (D_k), es decir:

$$Q_1 = Q_0 + \sum_{j=1}^{|R|} D_j - D_k \tag{4.2}$$

Otra alternativa para realizar este tipo de modificación de la consulta en el modelo que estamos tratando es la aplicación del *método de Rocchio*, el cual normaliza los pesos de los términos de los documentos relevantes y no relevantes dividiéndolos por el cardinal de cada uno de los conjuntos, como se puede ver en la fórmula (4.3). Además, se puede ajustar la importancia que se le da a ambos tipos de documentos mediante los valores que se asignen a los parámetros β y γ. Salton y Buckley [SB90] determinaron que los mejores eran 0,75 y 0,25, respectivamente, limitando así la influencia de los términos que aparecen en documentos no relevantes.

$$Q_1 = Q_0 + \beta \sum_{j=1}^{|R|} \frac{D_j}{|R|} - \gamma \sum_{j=1}^{|\bar{R}|} \frac{D_j}{|\bar{R}|} \tag{4.3}$$

Estos mismos autores, en [SB90], y Harman, en [Har92], han realizado un estudio muy detallado sobre el rendimiento de estos métodos anteriores, de donde se deduce que el que mejor se comporta es *Ide Dec-Hi*, aunque los porcentajes de cambio dependen claramente de la colección que se utilice.

En referencia a esto último, nos queda para completar la introducción a la realimentación de relevancia que aquí estamos haciendo, exponer cómo se determina la calidad de un método de realimentación, o lo que es lo mismo, medir su rendimiento. Como ya comentamos en el capítulo 1, las medidas de exhaustividad y precisión son las que se utilizan para determinar el rendimiento de un S.R.I., mediante el cálculo de la precisión para los once puntos estándar de exhaustividad.

Ahora bien, imaginemos la situación en la que un usuario emite sus juicios de relevancia con respecto a la salida obtenida por el S.R.I. a partir de una consulta inicial. El sistema de recuperación generará una nueva consulta a partir de dichos juicios y reiterará el proceso de recuperación con ella. Si nos fijamos en las posiciones que en esta segunda ordenación ocupan los documentos que previamente determinó el usuario como relevantes, observaremos cómo están todos ellos en unas posiciones muy altas. Esto implica que la correspondiente curva es mucho mejor que la asociada a la primera cuestión, pero en realidad la segunda ordenación aporta poca información nueva, pues esos documentos ya han sido inspeccionados y juzgados previamente.

Para evitar este problema, la evaluación del rendimiento de la realimentación de relevancia se suele llevar a cabo utilizando el método de la *colección residual* [CCR71], que consiste básicamente en eliminar de la colección los documentos que han sido juzgados en las listas ordenadas de documentos obtenidos con las sucesivas consultas. Por tanto, y considerando sólo un paso en la realimentación, el método de la colección residual funciona como sigue:

1. El S.R.I. genera una lista de documentos ordenada según la relevancia de cada uno de ellos con respecto a la consulta original.

2. Se presentan al usuario los n documentos más relevantes según el S.R.I. y aquél los inspecciona y emite sus juicios de relevancia.

3. El S.R.I. elimina de la colección los n documentos incluidos en estos juicios.

4. El S.R.I. genera una nueva consulta a partir de ellos y efectúa una recuperación con ella en la colección residual, obteniendo una nueva ordenación.

5. Se calcula la precisión para los once puntos estándar de precisión y la media de la precisión para los once o los tres puntos intermedios para la ordenación generada por la consulta inicial, pero eliminando los documentos que ha juzgado el usuraio.

6. Se calcula la curva exhaustividad – precisión para la ordenación de documentos de la segunda consulta y las dos medias de precisión.

7. Por último, se obtiene un porcentaje de cambio de la precisión media de la consulta modificada con respecto a la de la consulta inicial, consiguiendo una medida que refleja de manera más objetiva la mejora que se consigue con la aplicación de la realimentación.

4.2. Realimentación de relevancia en los principales modelos basados en redes bayesianas.

Centrémonos seguidamente en los dos principales modelos de recuperación basados en redes bayesianas: el modelo *Inference Network* de Croft y Turtle y el modelo *Belief Network Model* de Ribeiro y Reis. Para ellos también se han desarrollado métodos de realimentación de relevancia con los que introducen los juicios de relevancia expresados por un usuario al inspeccionar la lista de documentos devuelta por el S.R.I., mejorando finalmente el rendimiento del sistema recuperador. En esta sección vamos a explicar cómo se pone en práctica la realimentación de relevancia en estos dos modelos, la cual es conceptualmente muy parecida entre ellos: modificación de la topología de la red, mediante la inserción de nuevos nodos y/o arcos, y la actualización o nuevo cálculo, según el caso, de las matrices de probabilidad condicionada.

Comenzando con el modelo *Inference Network*, la realimentación básicamente se pone en práctica [HC93], en su vertiente de la expansión, mediante la creación de nuevos enlaces dirigidos que unen los términos de los documentos que se han juzgado como relevantes con el nodo de la consulta (es decir, colocando a los primeros como nuevos padres del nodo consulta), y posteriormente en la actualización de las matrices de probabilidad condicional a la luz de los nuevos arcos. En cuanto al repesado de términos, en este modelo se lleva a cabo recalculando de nuevo las matrices de probabilidad condicionada teniendo en cuenta la información de los documentos relevantes en la ordenación original. El proceso de recuperación concluye calculando la probabilidad de que la nueva consulta sea satisfecha dado que cada documento se ha instanciado individualmente como relevante.

En cuanto al modelo *Belief Network*, su método de realimentación se cimenta en dos operaciones fundamentales [Rei00]:

1. Modificación de la topología de la red. Cuando el usuario efectúa los juicios de relevancia, la red original se ve alterada de la siguiente manera (véase figura 4.1):

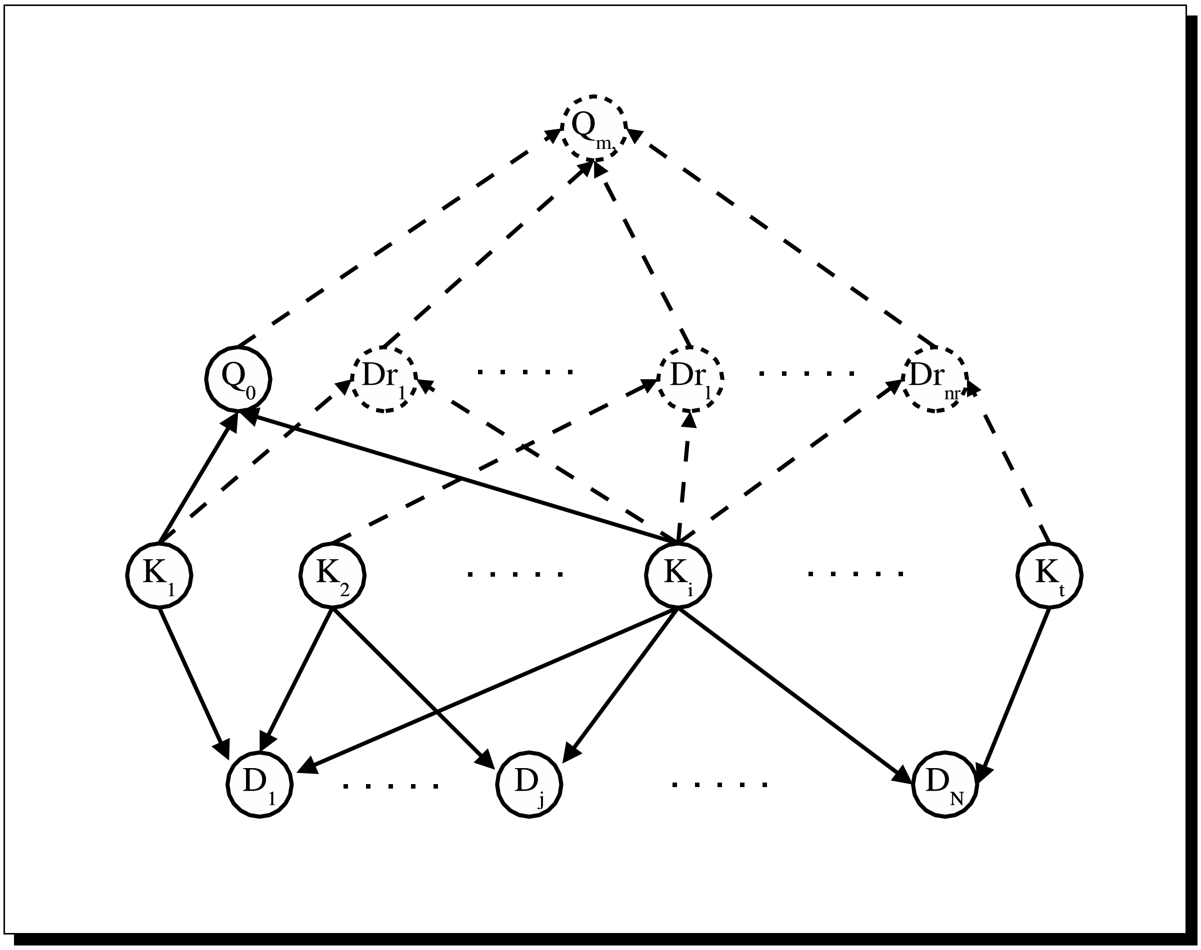


Figura 4.1.: Modificación de la topología de la red en el Belief Network Model al realizar la realimentación de relevancia.

- Se insertan tantos nuevos nodos de tipo documento Dr_i como nodos haya juzgado como relevantes el usuario.
- Se añade también un nuevo nodo de tipo consulta Q_m, enlazándolo a los nodos que representan los documentos relevantes y al nodo asociado con la consulta original, como sus padres.
- Se conectan los nodos documentos relevantes con los nodos términos asociados a los términos por los que fueron indexados, mediante arcos dirigidos desde los segundos a los primeros (en la figura, los arcos con trazo discontinuo).

2. Cálculo de las nuevas probabilidades originadas por la nueva topología. Las probabilidades de la red original permanecen intactas, aunque al haberse efectuado la adición de nuevos nodos y arcos, se tienen que calcular nuevas probabilidades condicionales: la consulta dado cada documento relevante, y los documentos relevantes dados los términos que contienen.

Finalmente, la red está dispuesta para soportar un nuevo proceso de propagación al instanciar los términos de la consulta original calculando, para cada documento de la colección, la probabilidad de que esté activo dada la nueva consulta.

La diferencia fundamental entre estos dos métodos radica en que el segundo usa unos nuevos nodos que representan a los documentos relevantes, mientras que el primero sólo inserta nuevos arcos. En ambos casos, cuando estas operaciones de actualización de la topología han finalizado, hay que realizar una estimación de las nuevas probabilidades condicionales que surgen o la reestimación de algunas ya existentes.

Como se podrá observar en la siguiente sección, nuestro método de realimentación de relevancia se basa en un enfoque completamente diferente, ya que no altera la red ni tiene que calcular nuevas matrices de probabilidad, pues su fundamento radica en la forma en que se instancian los nodos en la red documental.

4.3. Fundamentos sobre el método de realimentación desarrollado.

La metodología que aquí presentamos para ser aplicada al modelo de recuperación basado en redes bayesianas expuesto en el capítulo anterior, se fundamenta en el uso de *evidencias parciales*, que ya introdujimos en el capítulo 1 y utilizamos en el capítulo 3 para instanciar los términos de una consulta.

Cuando un usuario evalúa la lista de documentos devuelta por el S.R.I., éste puede obtener nuevas evidencias sobre la relevancia de los términos que indexan el conjunto de documentos juzgados con respecto a la necesidad de información del usuario. Así, si un término T_i indexa un documento que ha sido calificado como relevante, la creencia en la aseveración "el término T_i es relevante" se debería incrementar. Por el contrario, si este término apareciera en un documento no relevante, sería, en este caso, la creencia en "el término T_i no es relevante" la que se incrementaría. Decimos, pues, que existe una evidencia parcial en el conjunto de términos que ocurren en documentos relevantes y no relevantes, pues sólo con los juicios de relevancia sobre los documentos no podemos estar completamente seguros sobre la veracidad o no de las aseveraciones anteriores. Además, esta nueva información debe ser combinada con la evidencia apoyada por la consulta original, con objeto de formular una nueva consulta.

Supongamos que después de analizar un documento tuviéramos una creencia en que el término sea relevante diez veces superior al hecho de que no lo sea. Esta información se puede expresar por medio de un cociente de valores de verosimilitud, es decir,

$$P(\ Observación\ |\bar{t}_i) : P(\ Observación\ |t_i) = 1 : 10$$

donde $P(Observación|t_i)$ representa la probabilidad de obtener una observación dado que sabemos que el término T_i es relevante (análogamente para $P(Observación|\bar{t}_i)$).

Planteémosnos seguidamente la forma de incluir esta información en la red documental, para lo cual utilizaremos el concepto de *evidencia parcial*: si para un nodo X, término o documento, se encuentra una nueva evidencia apoyando su relevancia o irrelevancia, el nodo recibirá un mensaje, que notaremos $\lambda(X)$, que albergará el par de verosimilitudes siguientes:

$$\lambda(X) = a \cdot (p(\textit{ Observación } |\bar{x}), p(\textit{ Observación } |x))$$

siendo a una constante de normalización, que permitirá que los vectores λ puedan tomar cualquier valor. En nuestro caso, los vectores se normalizarán dividiendo por la verosimilitudes $p(\textit{ Observación } |x)$. Veamos un ejemplo de vectores equivalentes: dado el vector $\lambda_1(X) = a_1 \cdot (1,10)$, éste es equivalente a $\lambda(X) = a_2 \cdot (\frac{1}{11}, \frac{10}{11})$ y a $\lambda(X) = a_3 \cdot (0{,}1, 1)$, representando los tres vectores la misma evidencia para X.

En realidad, cuando existen evidencias parciales sobre un nodo X de la red, se crea un nodo imaginario y se conecta como hijo de X, que representa la evidencia sobre esta variable. En el proceso de inferencia, este nuevo nodo le enviará un mensaje λ a X, que será combinado con toda la información que reciba X de sus padres y de sus hijos.

Las creencias sobre la relevancia o no de un término pueden ser cuantificadas numéricamente e insertadas en la red bayesiana por medio de los vectores λ, con lo cual, y teniendo en cuenta este hecho, el proceso que sigue nuestro método de realimentación es conceptualmente bien sencillo: una vez que el usuario ha efectuado los juicios de relevancia correspondientes, el S.R.I. recorre todos los documentos relevantes y no relevantes, y para cada uno de los términos que aparecen en ellos calcula un vector λ que, según el caso, puede representar una combinación de varios vectores, o bien, un único vector que resuma la información obtenida a partir de la ordenación de documentos. Posteriormente, esos vectores se usan como evidencias parciales, pasando a propagarse por toda la subred de términos.

Los dos métodos desarrollados para realizar la realimentación en el modelo de red bayesiana documental se basan en el uso subyacente de evidencias parciales, aunque corresponden a dos enfoques conceptualmente diferenciados. El primero de ellos, el *método de realimentación basado en los términos*, establece que mediante el enjuiciamiento de los documentos recuperados el usuario puede obtener nuevas evidencias de manera exclusiva para los términos, momento en el cual aparece la figura de las evidencias parciales. Por otro lado, el *método de realimentación basado en los documentos* parte del hecho de que las evidencias obtenidas a partir de la inspección de la ordenación de documentos se centran en los documentos, y no en los términos, de tal manera que sólo se instancian los propios documentos que han sido juzgados como relevantes o no.

Dichos métodos serán expuestos según la línea clásica utilizada en los trabajos realizados en este campo, como por ejemplo el de Harman [Har92]. Así, en primer lugar, y para cada uno de ellos, comenzaremos tratando las técnicas de repesado de consultas, pasaremos seguidamente a la expansión y posteriormente a la combinación de ambas, para finalizar con el estudio del impacto de una selección de términos en el rendimiento de la realimentación, mostrando paralelamente el rendimiento alcanzado con la colección CACM en cada uno de estos apartados anteriores. Finalmente, y para determinar con más precisión la calidad de los métodos propuestos, expondremos los resultados de la experimentación con el resto de colecciones.

Antes de comenzar, debemos hacer varias consideraciones generales que serán tenidas en cuenta a la hora de poner en práctica los métodos de realimentación que vamos a exponer seguidamente:

- Nuestro S.R.I. devuelve al usuario un total de quince documentos del total de la colección.
- Se tienen en cuenta tanto las consultas en las que se recuperan documentos relevantes como en las que no. Hacemos esta puntualización debido a que en muchos trabajos estas consultas no se tratan, aumentando así el rendimiento de la realimentación. Sólo se descartan, por razones obvias, las consultas que recuperan todos los documentos relevantes.
- Como trabajamos con colecciones documentales estándar que ya traen su propio conjunto de consultas, la evaluación de la realimentación se efectúa de la siguiente manera: se suma la media de la precisión de los tres puntos de exhaustividad obtenida en todas las consultas originales sobre la colección residual. Seguidamente, se realiza la misma operación para los valores conseguidos al llevar a cabo la realimentación. Por último, se calcula el porcentaje de cambio de la segunda cantidad con respecto a la primera.
- Sólo efectuamos una iteración de realimentación, es decir, la consulta original más la correspondiente consulta modificada.
- En cuanto al tipo de red con el que se han llevado a cabo los experimentos de realimentación, habría tres alternativas posibles: elegir una red y una combinación de argumentos con los que se alcance un rendimiento muy bueno en cada colección, una efectividad recuperadora baja, por el contrario, y finalmente, media. Si se toma la primera opción, la calidad del método de realimentación puede verse diluida en las buenas condiciones recuperadoras de la red y sus argumentos, con lo que no se aprecia realmente su efectividad. Algo análogo, pero negativamente, ocurre si se toma la opción segunda: la red no ofrece una base buena para modificar la consulta y los resultados obtenidos pueden ser bastante mediocres. Por otro lado, con una red cuyo comportamiento esté entre ambas, creemos que se permite observar más claramente la calidad recuperadora de los métodos de realimentación. Por esta razón, hemos elegido la red aumentada, con probabilidades a priori estimadas según *pp2* y probabilidades condicionadas calculadas aplicando *pc-mv*. La propagación se ha efectuado sin evidencias parciales, sin intervención de los pesos de la consulta y usando la función de probabilidad *fp10*. En la tabla 4.1 se pueden observar los porcentajes de cambio para las cinco colecciones con respecto a SMART y los resultados de la combinación de red documental y parámetros anteriormente descrita.
- Los porcentajes de cambio que se ofrecen corresponden a la media de los tres puntos intermedios de exhaustividad, es decir, 0,2, 0,5 y 0,8. Se ha elegido la media de estos tres puntos por establecer una compatibilidad con respecto a los resultados ofrecidos en [HC93] para la realimentación en el modelo *Inference Network*.

Colección	Media 3 puntos SMART	Media 3 puntos red experimentos	% de cambio
ADI	0.4848	0.3627	-25.2
CACM	0.3683	0.3438	-6.7
CISI	0.2272	0.1656	-27.1
CRANFIELD	0.4269	0.4177	-2.1
MEDLARS	0.5527	0.6304	14.1

Cuadro 4.1.: Medias de tres puntos de exhaustividad para SMART y la red documental usada en la experimentación de realimentación.

	$\bar{t}$	t	
$\bar{r}$	$n_{\bar{r}\bar{t}}$	$n_{\bar{r}t}$	$n_{\bar{r}}$
r	$n_{r\bar{t}}$	n_{rt}	n_r
	$n_{\bar{t}}$	n_t	$\mid J \mid$

Cuadro 4.2.: Frecuencia de aparición de un término en documentos relevantes y no relevantes.

- En cuanto a notación se refiere, la tabla 4.2 muestra la frecuencia de aparición o no de un término cualquiera, T, en el conjunto de los documentos calificados por el usuario como relevantes o no relevantes (este conjunto se representará como J) y pertenecientes a la ordenación de los $\mid J \mid$ primeros documentos recuperados por el S.R.I. El significado de los diferentes elementos que la componen es el siguiente:

    - r significa "el documento es relevante".
    - $\bar{r}$, "el documentos no es relevante".
    - n_r es el número de documentos relevantes recuperados.
    - n_t es el número de ocasiones en el que el término T se ha observado entre los $\mid J \mid$ primeros documentos.
    - $n_{\bar{t}}$ es el número de documentos de entre los $\mid J \mid$ primeros en los que no aparece dicho término.
    - n_{rt} es el número de documentos relevantes que han sido indexados por T.
    - $n_{\bar{r}t}$ es el número de documentos no relevantes indexados por T.
    - $n_{r\bar{t}}$ es el número de documentos relevantes en los que no aparece T.
    - $n_{\bar{r}\bar{t}}$ es el número de documentos no relevantes donde no aparece el término T.

4.4. Método de realimentación de relevancia basado en los términos.

La idea básica que yace en él es que para cada término de los documentos que han sido calificados como relevantes o no relevantes, se creará un nodo imaginario, y éste calculará a partir de la información de la lista ordenada de documentos y de la relativa a la colección un único vector λ que será el que cuantifique la evidencia parcial. Dependiendo del tipo de término, este vector λ se calculará de una u otra forma, como a continuación veremos.

4.4.1. Repesado de la consulta.

En primer lugar, sólo vamos a tratar el problema del repesado de la consulta, es decir, cómo se modifican los pesos de los términos de la consulta original a la luz de los juicios de relevancia efectuados por el usuario. En nuestro modelo, los términos de la consulta original se instancian al vector $(0,1)$, indicando así que han sido observados como totalmente relevantes. El resto de términos se instancia a $(1,1)$, hecho que es equivalente a no instanciarlos. De esta forma, teniendo en cuenta la herramienta de que disponemos, la única manera de repesar los términos originales pasará por el cálculo de un nuevo vector λ para cada uno de ellos.

Podríamos calificar esta técnica de repesado como un tanto negativa, ya que la total seguridad en la relevancia de un término se refleja con el vector $(0,1)$, vector al que ya han sido instanciados los términos de la consulta original, y por tanto, no se puede premiar más a un término original que se considere muy bueno. Por el contrario, a aquellos términos que no estén realizando una buena labor, se les puede quitar importancia en la nueva consulta mediante el cálculo de un vector λ cercano a $(1,1)$, con lo cual se puede llegar a sacar estos términos de la consulta.

Atendiendo a la aparición de los términos originales (aquéllos que figuran en la consulta inicial) en los documentos que han sido juzgados como relevantes y no relevantes por el usuario, los clasificaremos a efectos de notación en los tres siguientes conjuntos disjuntos:

1. *Términos de la consulta positivos (tc+)*: aquellos que pertenecen a la consulta original y sólo aparecen en documentos relevantes.
2. *Términos de la consulta negativos (tc-)*: los que están incluidos en la consulta original que sólo se encuentran en documentos juzgados como no relevantes.
3. *Términos de la consulta neutros (tc=)*: miembros de la consulta original que están tanto en documentos relevantes como en no relevantes.

Teniendo en cuenta esta clasificación, seguidamente vamos a explicar la forma de cálculo de los vectores λ para cada uno de los tres grupos anteriores:

1. *Términos de la consulta positivos.*

   Al ser términos que sólo aparecen en documentos relevantes, parece razonable que se sigan considerando totalmente relevantes, ya que están realizando correctamente la misión para la que fueron incluidos en la consulta. Por tanto, el vector λ no se modificará, es decir, será $\lambda(T_i) = (0, 1)$.

2. *Términos de la consulta negativos.*

   Estos términos sólo se encuentran en los documentos no relevantes, y por esta razón deben ser penalizados, es decir, calificarlos como parcialmente relevantes. Esta idea se puede poner en práctica elevando el componente negativo del vector λ, es decir, calculando un vector de la forma: $\lambda(T_i) = (\gamma, 1)$, con $0 < \gamma < 1$. Hay que ser bastante cuidadoso con el valor γ que se ponga en el componente negativo del vector, ya que si éste es próximo a 1, se corre el riesgo de sacar de la consulta al término, ya que el vector será muy parecido a $(1, 1)$, el cual expresa nuestra máxima incertidumbre. Por otro lado, si γ está muy cercano a 0, los resultados no se verán prácticamente afectados, debido a que se está instanciando casi a $(0, 1)$ de nuevo.

   Nuestra propuesta para repesar este tipo de términos pasa por dos alternativas:

   a) Utilizar un valor fijo γ para todos los términos de la consulta negativos, o

   b) hacer que el valor de γ dependa del número de documentos no relevantes en los que aparezca. De esta manera, cuanto mayor sea el número de éstos, γ será también mayor. Atendiendo a esta condición, el vector λ para un término $T_i \in tc-$ viene dado por:

   $$\lambda(T_i) = (1 - \frac{1}{n_{\bar{r}t} + 1}, 1) \tag{4.4}$$

   Esta forma de repesar los términos de la consulta negativos se denomina *rc1*, y asegura que $0{,}5 \leq \lambda < 1$, con lo cual la influencia del término en la consulta original se ve debilitada en la nueva.

3. *Términos de la consulta neutros.*

   En este caso estamos tratando términos de la consulta que aparecen en ambos tipos de documentos. Hay dos opciones principales:

   a) Como son términos que aparecen en documentos relevantes, instanciarlos a $(0, 1)$.

   b) La segunda opción pasa por tener en cuenta el número de documentos de los dos tipos donde aparecen, configurando así el método de repesado *rc2*. Los términos de la consulta neutros se instanciarán a un vector distinto según su distribución en los conjuntos de documentos relevantes y no relevantes, dados por las siguientes expresiones:

Exp.	tc+	tc-	tc=	% de cambio
1	$(0,1)$	$(0{,}2,1)$	$(0,1)$	28,23
2	$(0,1)$	$(0{,}5,1)$	$(0,1)$	29,83
3	$(0,1)$	$(0{,}8,1)$	$(0,1)$	30,05
4	$(0,1)$	$rc1$	$(0,1)$	29,90
5	$(0,1)$	$rc1$	$rc2$	$-52{,}08$

Cuadro 4.3.: Resultados del repesado de la consulta original con la colección CACM.

- Si $n_{rt} = n_{\bar{r}t} \Rightarrow \lambda(T_i) = (0{,}5, 1)$
  El término se penaliza elevando la verosimilitud de la no relevancia hasta 0,5.
- Si $n_{rt} > n_{\bar{r}t} \Rightarrow \lambda(T_i) = (\frac{1}{n_{rt}-n_{\bar{r}t}+1}, 1)$
  La componente negativa del vector está limitada inferiormente por 0 y superiormente por 0,5, y además, se hace más pequeña cuando la diferencia del número de documentos relevantes con respecto a los no relevantes crece.
- Si $n_{rt} < n_{\bar{r}t} \Rightarrow \lambda(T_i) = (1 - \frac{1}{n_{\bar{r}t}-n_{rt}+1}, 1)$
  En este caso, la componente negativa tomará como máximo el valor 1 y 0,5 como mínimo, y se incrementará conforme la diferencia entre el número de documentos no relevantes donde aparezca y relevantes se incremente.

Con el método *rc2* se pretende acercar $(1,1)$ los términos originales que aparecen en muchos documentos no relevantes en relación al número de relevantes donde se han visto. Por otro lado, aquéllos cuya tendencia sea la de figurar en más documentos relevantes que no relevantes, penalizarlos algo, separándolos de $(0,1)$ (no son positivos puros), pero no excesivamente, con objeto de que sigan realizando correctamente su labor.

Los resultados que se ofrecen en la tabla 4.3 corresponden a una primera fase de experimentación con la colección CACM, en la que se ponen a prueba las diferentes técnicas de repesar que se acaban de comentar.

Los tres primeros experimentos se han realizado modificando sólo los términos de la consulta negativos, mediante una instanciación de la forma $(\gamma, 1)$, variando el valor del parámetro. Tanto los términos de la consulta positivos como los neutros se han instanciado a $(0,1)$. Así, para observar cómo cambia el rendimiento con respecto al valor asignado a γ, se han llevado a cabo pruebas para los valores $0,2$, $0,5$ y $0,8$. Se aprecia a la luz de los resultados que el porcentaje de cambio se incrementa con respecto al aumento de γ, aunque dicho cambio es poco significativo. La segunda alternativa es la aplicación del método *rc1*, que se corresponde con el experimento 4 de la tabla, y que ofrece un porcentaje de cambio muy aproximado a los que se consiguen con las tres primeras pruebas, hecho que evita tener que encontrar, para cada colección, el valor para el cual el rendimiento es mayor. Por último, el quinto experimento aplica el método *rc2* a los términos de la consulta neutros, en lugar de instanciarlos a $(0,1)$, en conjunción con el método *rc1* para los negativos. Del resultado mostrado en esta tabla se puede deducir que es preferible dejar los términos de la consulta neutros instanciados a $(0,1)$, ya que el hecho de quitarlos dependiendo de su distribución en documentos relevantes o no es realmente perjudicial.

Así las cosas, y atendiendo a los resultados empíricos generados con la colección CACM, la mejor técnica para repesar los términos de la consulta en nuestro modelo pasa por calcular el vector λ correspondiente para los términos negativos, según el método *rc1*, e instanciar el resto, los positivos y los neutros, a $(0,1)$.

4.4.2. Expansión de la consulta.

El objetivo de esta sección es determinar el impacto que produce la adición de nuevos términos a la consulta original, dejando los términos de esta consulta intactos, es decir, instanciados a $(0,1)$.

Como hicimos en la subvención anterior con los términos de la consulta, los términos susceptibles de ser añadidos a la consulta original para formar la consulta expandida, los cuales serán denominados *términos de expansión*, se clasifican en los tres siguientes grupos:

1. *Términos de expansión negativos (te-)*: los que sólo se encuentran en documentos juzgados como no relevantes sin pertenecer a la consulta original.
2. *Términos de expansión positivos (te+)*: aquéllos que sólo aparecen en documentos relevantes y no están incluidos en la consulta original.
3. *Términos de expansión neutros (te=)*: los que sin estar en la consulta original están tanto en documentos relevantes como en no relevantes.

Pasemos seguidamente a exponer las tácticas de ponderación aplicables a cada una de estas clases de términos de expansión:

1. *Términos de expansión negativos (te-)*: se instanciarán a "no relevante", es decir, $\lambda(T_i) = (1,0)$, ya que sólo introducen en lista ordenada de documentos devuelta por el S.R.I. documentos no relevantes.
2. *Términos de expansión positivos (te+)*: la evidencia a favor o en contra de la relevancia de un término de expansión positivo se basa en la tabla de contingencia 4.2 que se calcula para cada término. Veamos las diferentes aproximaciones desarrolladas en este enfoque:
   - *t1*: las verosimilitudes sobre la relevancia o no de los términos aportadas por los documentos relevantes vienen dadas por las siguientes expresiones:

$$p(r|t) = \frac{n_{rt}}{n_t} \text{ y } p(r|\bar{t}) = \frac{n_{r\bar{t}}}{n_{\bar{t}}} \quad (4.5)$$

O lo que es lo mismo, la probabilidad de que un documento que contenga al término de expansión sea relevante, y la probabilidad de relevancia de un documento que no contenga a dicho término.

Teniendo en cuenta estas verosimilitudes, el vector λ que se calcula para ese término queda como sigue:

$$\lambda(T_i) = (p(r|\bar{t}), p(r|t)) = (\frac{p(r|\bar{t})}{p(r|t)}, 1) \tag{4.6}$$

Como se observa en la segunda igualdad, el segundo componente del vector se ha asignado a 1, mientras que el primero se puede expresar como el cociente de $\lambda(\bar{t})$ entre $\lambda(t)$, ya que lo que verdaderamente interesa no es la magnitud de estas verosimilitudes, sino la relación entre ellas, es decir, la proporción.

- *t2*: Debido a que estamos tratando con un número reducido de documentos (los $\mid J \mid$ primeros de la ordenación), las frecuencias de la tabla 4.2 suelen ser muy bajas, por lo que la estimación de las probabilidades expuestas en (4.5) será de una baja calidad. Para solucionar este problema, volvemos a echar mano del *estimador bayesiano* [Goo65], con objeto de mejorar la estimación. Así, las verosimilitudes expuestas en (4.5) pasarán a ser:

$$p(r|t) = \frac{n_{rt} + s_t \frac{n_r}{|J|}}{n_t + s_t} \quad \text{y} \quad p(r|\bar{t}) = \frac{n_{r\bar{t}} + s_{\bar{t}} \frac{n_r}{|J|}}{n_{\bar{t}} + s_{\bar{t}}} \tag{4.7}$$

  En esta expresión, los parámetros s_t y $s_{\bar{t}}$ representan el tamaño muestral equivalente, aunque se va a considerar el mismo para ambas verosimilitudes, es decir, $s_t = s_{\bar{t}}$.

- *t3*: Aunque en la aproximación *t2* ambos parámetros van a tomar el mismo valor, puede darse el caso de que las verosimilitudes de la expresión (4.7) se estimen con un número distinto de datos cada una. Si la primera, por ejemplo, se estima con menos datos que la segunda, su fiabilidad será también menor que la de la otra, lo que implica a su vez que el parámetro s_t deberá tomar un valor mayor que $s_{\bar{t}}$ para corregir el problema. Por tanto, y según este razonamiento, los parámetros deberán ser inversamente proporcionales al número de datos con los que se haga la estimación. El caso extremo será aquel en el que $n_t = 15$ o $n_{\bar{t}} = 15$, en donde el parámetro correspondiente será 0, ya que se están utilizando todos los datos posibles. Basándonos en esta idea, el método *t3* utiliza la expresión (4.7) con los siguientes valores para los parámetros:

$$s_t = 15 - n_t = n_{\bar{t}} \text{ y } s_{\bar{t}} = 15 - n_{\bar{t}} = n_t \tag{4.8}$$

- *t4*: Este último enfoque para pesar los términos de expansión positivos se basa también en el uso de valores diferentes para ambos parámetros, pero en este caso se tiene en cuenta el número total de documentos en los que aparece y no aparece un término. Así, por ejemplo, si un término sólo indexa tres documentos y éstos han sido recuperados entre los quince primeros, no hará falta aplicar el estimador bayesiano, por lo que s_t deberá ser 0, ya que se están utilizando todos los datos disponibles. El caso contrario es el que se daría cuando el término indexa una gran cantidad de documentos, por ejemplo, doscientos, y sólo se ha recuperado uno. En este caso, el parámetro s_t deberá ser bastante alto.

Esta idea puede llevarse a la práctica dividiendo la frecuencia de un término en la colección, tf_i, y el número de veces que ha sido observado en la ordenación de los $| J |$ primeros, n_t. Teniendo en cuenta que este cociente puede ser muy alto en algunas ocasiones, se utilizan logaritmos para suavizar los valores que se obtienen y, consecuentemente, y atendiendo a las características especiales de esta función, se debe sumar la unidad para evitar indefiniciones. Por tanto, los parámetros en el enfoque *t4* se calculan como sigue:

$$s_t = \frac{log(tf_i + 1)}{log(n_t + 1)} + 1 \text{ y } s_{\bar{t}} = \frac{log(N - tf_i + 1)}{log(n_{\bar{t}} + 1)} + 1 \tag{4.9}$$

De nuevo, para obtener la expresión final del vector λ según este enfoque, hay que sustituir la expresión (4.9) en la (4.7).

3. *Términos de expansión neutros (te=)*: en este caso tenemos dos alternativas posibles:

   - Instanciarlos a $(1,1)$, con lo cual no se tienen en cuenta, o
   - *t5*: instanciarlos a un valor diferente que dependa de su distribución en documentos relevantes y no relevantes. Este enfoque de ponderación de los términos de expansión neutros los instancia a $(1,1)$ si aparecen en el mismo número de documentos relevantes que no relevantes, a $(1,0)$ si el número de documentos no relevantes donde aparece el término es mayor que el número de relevantes, y en el caso contrario, se instancia al vector obtenido por uno de los métodos expuestos anteriormente al tratar los términos de expansión positivos, quedando finalmente el método *t4*:

$$\lambda(T_i) = \begin{cases} (1,1) & \text{si} \quad n_{rt} = n_{\bar{r}t} \\ (\gamma,1) & \text{si} \quad n_{rt} > n_{\bar{r}t} \\ (1,0) & \text{si} \quad n_{rt} < n_{\bar{r}t} \end{cases} \tag{4.10}$$

Una vez expuestas todas las técnicas que se han desarrollado para la ponderación de los términos de expansión, vamos a pasar a estudiar su rendimiento al aplicarlas a la colección CACM, resultados que figuran en la tabla 4.4. En ella, la primera columna (*Exp.*) indica el número de experimento; la segunda, *tc*, establece el vector al cual han sido instanciados los términos pertenecientes a la consulta original, que en este caso, y al estar presentando experimentos sobre expansión de consultas, no se han modificado, instanciándose, por tanto, a $(0,1)$ todos ellos. Las columnas tercera, cuarta y quinta, contienen los vectores o los métodos que se han aplicado en cada experimento a los términos de expansión negativos, neutros y positivos,

Experimento	tc	te-	te=	te+	% de cambio
1	$(0,1)$	$(1,1)$	$(1,1)$	$(0,1)$	$-69{,}42$
2	$(0,1)$	$(1,0)$	$(1,1)$	$(0,1)$	$-59{,}73$
3	$(0,1)$	$(1,0)$	$(1,1)$	$t1$	3,0
4	$(0,1)$	$(1,0)$	$(1,1)$	$t2 : s_t, s_{\bar{t}} = 5$	46,36
5	$(0,1)$	$(1,0)$	$(1,1)$	$t2 : s_t, s_{\bar{t}} = n_r$	48,40
6	$(0,1)$	$(1,0)$	$(1,1)$	$t3$	38,76
7	$(0,1)$	$(1,0)$	$(1,1)$	$t4$	59,45
8	$(0,1)$	$(1,1)$	$(1,1)$	$t4$	12,19
9	$(0,1)$	$(1,0)$	t5	$t4$	45,11

Cuadro 4.4.: Experimentos sobre expansión de consultas en la colección CACM con el método basado en los términos.

respectivamente. La última columna (*% de cambio*) muestra el porcentaje de cambio obtenido en cada ensayo.

El primer experimento de la tabla 4.4 muestra cómo el expandir la consulta original con todos los términos de expansión positivos instanciados a $(0,1)$ no es una buena idea, ya que la nueva consulta pierde el sentido primario al añadir una gran cantidad de términos en iguales condiciones que los que figuran inicialmente en ella. El segundo intenta medir el impacto que tiene añadir los términos de expansión negativos en el rendimiento de la consulta. Se observa cómo éste mejora, aunque sigue siendo bastante malo debido a la fuerza que ejercen los términos positivos, lo cual indica que puede ser buena idea instanciar negativamente. El tercero comienza a probar los diferentes enfoques desarrollados para los términos de expansión positivos. Fijando la instanciación de los negativos a $(1,0)$ y de los neutros a $(1,1)$, este ensayo intenta determinar la calidad del método *t1*. Aunque el porcentaje de cambio es muy bajo, ya es positivo y considerablemente mejor que el de los anteriores, lo que prueba que es bueno utilizar un vector λ distinto a $(0,1)$. Donde ya empiezan a establecerse diferencias significativas en el rendimiento es en los cuatro siguientes experimentos. El cuarto usa el método *t2* con $s_t = s_{\bar{t}} = 5$, mientras que en el quinto quedan igualados al número de documentos relevantes recuperados en cada consulta. Los porcentajes de cambio de estas dos pruebas son muy parecidos, aunque ligeramente mayor para el segundo. Con respecto a las tentativas realizadas con *t3* y *t4*, el primero ofrece el rendimiento peor de los cuatro, al contrario que ocurre con el segundo, que consigue el mayor porcentaje de cambio de todos los métodos. A la luz de estos datos, podemos claramente deducir que es vital el uso de un estimador bayesiano en esta familia de métodos. Para reafirmar la idea de usar la instanciación negativa, el penúltimo experimento no los incluye en la consulta y, como se puede apreciar, el porcentaje de cambio disminuye en casi un 47 %, demostrando así su utilidad. Con el noveno y último experimento deseamos determinar el rendimiento del método *t5* para términos de expansión neutros. El método que utiliza es *t4*, ofreciendo un descenso considerable con respecto al ensayo séptimo, donde no se tenían en cuenta.

Exp.	te+	% de cambio
1	$t2 : s_t = s_{\bar{t}} = 5$	65,25
2	$t2 : s_t = s_{\bar{t}} = n_r$	67,11
3	$t3$	58,30
4	$t4$	72,08

Cuadro 4.5.: Experimentos sobre repesado y expansión de consultas en la colección CACM para el método basado en los términos.

4.4.3. Repesado de la consulta original más expansión.

En esta sección vamos a mostrar cómo repesando los términos de la consulta original y, a la vez, añadiéndoles nuevos términos, el rendimiento de la realimentación de relevancia aumenta considerablemente.

En los experimentos que hemos llevado a cabo, los términos de la consulta positivos se han instanciado a $(0,1)$, al igual que los neutros. Los negativos lo hacen según la técnica de repesado *rc1*. Por otro lado, y en lo que se refiere a la expansión, los términos de expansión negativos se instancian a $(1,0)$, los neutros no se tienen en cuenta (o lo que es lo mismo, su vector λ toma el valor $(1,1)$, y por último, los positivos tomarán diferentes valores dependiendo de la técnica de ponderación usada.

La tabla 4.5 muestra los resultados conseguidos combinando el repesado con las técnicas de expansión presentadas. En ella notamos cómo los porcentajes de cambio crecen considerablemente con respecto a los obtenidos separadamente en la tablas 4.4 y 4.3, demostrando así la conveniencia de aplicar el repesado y la expansión simultáneamente. La técnica de expansión que alcanza un porcentaje de cambio mayor es *t4*, la que rinde peor es *t3*, y *t2*, con sus dos valores diferentes de los parámetros, presenta una conducta entre las dos anteriores. No se ha efectuado ningún experimento con la técnica *t1*, ya que sólo con la expansión ofrecía unos resultados muy pobres, y la mejora que se podría obtener añadiéndole el repesado seguiría siendo bastante pobre comparada con el resto de miembros de su familia de técnicas de expansión.

4.4.4. Selección de términos.

Por último, para acabar el estudio de los diferentes componentes que están disponibles para ser usados en la realimentación de relevancia, nos queda determinar el número de términos de expansión que se añaden a la consulta y con el que se conseguirá un porcentaje de cambio mayor.

De los dos tipos de términos de expansión que se añaden a la consulta original, los negativos y los positivos, vamos a realizar la selección de términos sólo con los segundos, ya que adoptamos el criterio de incluir cualquier término que sólo apareciera en documentos de la ordenación juzgados como no relevantes. La selección se pone en práctica ordenando todos los términos de

Exp.	Número de términos seleccionados	% de cambio
1	10	64,30
2	20	62,43
3	30	61,96
4	Todos	72,08

Cuadro 4.6.: Resultados de la selección de términos con la colección CACM y la técnica de expansión *t4*.

tc+	tc-	tc=	te+	te-	te=
$(0,1)$	*rc1*	$(0,1)$	*t1, t2, t3 y t4*	$(1,0)$	$(1,1)$

Cuadro 4.7.: Conjunto de experimentos realizados con ADI, CISI, CRANFIELD y MEDLARS.

manera creciente de acuerdo con el primer componente de su vector λ, es decir, $\lambda(\bar{t}_i)$, ya que cuanto menor sea este valor, más interesa el término.

Los ensayos efectuados utilizan el repesado $rc1$ más la técnica de expansión *t4* (ya que con ella se han obtenido los mejores rendimientos de la realimentación), pero seleccionando cada vez 10, 20 ó 30 términos de expansión positivos, resultados presentados en la tabla 4.6.

Claramente se puede advertir cómo cuantos más términos de expansión positivos se introducen, menor es el porcentaje de cambio, aunque al final la tendencia cambia, y añadiendo todos los términos de este tipo, el porcentaje de cambio se incrementa sustancialmente en casi diez puntos porcentuales. Esta conducta es justo al contrario que la comentada por Harman [Har92] en sus experimentos, en los que al reducir el número de términos expandidos se mejoraba el rendimiento. De hecho, demuestra empíricamente que la mejor cantidad es quince.

4.4.5. Experimentación con las colecciones ADI, CISI, CRANFIELD y MEDLARS.

Una vez realizada la presentación del método de realimentación desarrollado y estudiado de manera detallada su comportamiento con la colección CACM, vamos a pasar a determinar qué rendimiento se obtiene cuando se aplica a otras colecciones como son ADI, CISI, CRANFIELD y MEDLARS.

Con estas colecciones sólo se mostrarán los resultados correspondientes a los experimentos que figuran en la tabla 4.7, los cuales están basados en el repesado de la consulta original más la expansión de la misma con todos los términos de expansión negativos y positivos.

Comencemos comentando los resultados (tabla 4.8) para la colección ADI, cuyo comportamiento es muy parecido al conseguido por la colección CACM: el mejor rendimiento se alcan-

Experimentos	te+	ADI	CACM	CISI	CRANFIELD	MEDLARS
1	$t1$	104,2	6,33	8,2	99,1	$-32,6$
2	$t2 : s_t = s_{\bar{t}} = 5$	71,7	65,3	41,0	98,1	10,1
3	$t2 : s_t = s_{\bar{t}} = n_r$	95,0	67,1	42,9	99,9	7,9
4	$t3$	68,6	58,3	42,7	94,6	8,4
5	$t4$	105,0	72,1	39,0	107,7	12,4

Cuadro 4.8.: Resultados obtenidos con el repesado y expansión con el método basado en los términos para ADI, CISI, CRANFIELD y MEDLARS.

za aplicando *t4*, aunque la técnica *t1*, que en la CACM consigue un porcentaje de cambio muy pobre, llega al segundo valor más alto, aunque muy próximo al primero. Las dos variaciones de *t2* tienen un comportamiento calificable como bueno, aunque el caso en el que $s_t = s_{\bar{t}} = n_r$ destaca con respecto al otro experimento de la misma técnica. Por último, el peor rendimiento lo ofrece *t3*, aunque está cercano al 70 %.

En cuanto a CISI, todos los porcentajes de cambio son muy parecidos entre sí, excepto el que se alcanza al aplicar *t1*, el cual es bastante bajo. En este caso, es *t2* la técnica que ofrece un rendimiento mayor. En general, el rendimiento de las técnicas en esta colección es pobre, ya que no llega a un 43 %.

CRANFIELD alcanza, al igual que ADI, porcentajes de cambio bastante altos. Los tres primeros experimentos están muy próximos entre sí, rondando el 100 %, aunque es *t4* la técnica que se destaca sobre las demás casi en 7 puntos.

Por último, MEDLARS es la colección que tiene los porcentajes de cambio más bajos con diferencia. La razón se puede deber a que el mecanismo de recuperación aplicado a la red con la que se han efectuado estos experimentos consigue una precisión media para los tres puntos de exhaustividad bastante alta. De nuevo es *t4* la mejor y *t1* la peor.

En general, parece que $t4$ es la técnica de expansión que se comporta mejor en casi todas las colecciones, seguida en la mayoría de las ocasiones por $t2$ en su versión en la que los dos parámetros son iguales a n_r.

4.5. Método de realimentación de relevancia basado en los documentos.

Como dijimos en secciones previas, la diferencia fundamental de este método que vamos a presentar a continuación y el basado en los términos, radica en que se instancian directamente los documentos, en lugar de los términos contenidos en ellos. Así, una nueva consulta donde se ha aplicado la realimentación quedará compuesta por la siguiente estructura:

$$Q_1 = (Q, d_1, d_2, \ldots, d_k, \bar{d}_{k+1}, \ldots, \bar{d}_{|J|})$$

donde $| J |$ es el número de documentos que el usuario ha juzgado y d_i y $\bar{d}_j$ indican que los documentos D_i y D_j se han calificado como relevante e irrelevante, respectivamente.

En esta situación, lo ideal es realizar una propagación en la red completa, instanciando nodos término y nodos documento a la vez. Debido a que la propagación exacta a través de toda la red es inviable, como ya expusimos en el capítulo anterior, debemos buscar una forma alternativa con la cual aproximar los cálculos de la propagación.

En el proceso de propagación, y como consecuencia de la instanciación, cada nodo documento le manda un mensaje λ a sus padres (el conjunto de nodos que representan a los términos que lo indexan), conteniendo la información de que ha sido juzgado como relevante o irrelevante. La idea básica de esta aproximación es simular la instanciación: calcular de manera aproximada los vectores λ que mandaría a sus padres, cálculos que se basarán en las probabilidades almacenadas en la red documental. Una vez que el nodo documento los ha calculado y enviado, los nodos términos que los reciben deberán combinarlos para conseguir el valor de la evidencia global.

El esquema de exposición de esta técnica va a ser totalmente análogo al seguido con el método basado en los términos, aunque para el caso del repesado de la consulta hay que hacer la siguiente puntualización: considerando que los términos pertenecientes a la consulta original se han instanciado a relevantes, nuestra creencia en su relevancia no se verá modificada mediante la instanciación de los documentos, razón por la cual se usará como técnica de repesado de la consulta la expuesta en la sección 4.4.1, es decir, *rc*1. Esta situación que acabamos de describir no se mantiene para los términos que no están en la consulta original y que indexan a los documentos observados, donde los mensajes recibidos de los documentos juzgados modifican la creencia sobre su relevancia. Es por esto que comenzaremos directamente la descripción del método por la expansión de la consulta, que equivaldrá a la instanciación de documentos.

4.5.1. Expansión de la consulta.

En primer lugar nos centraremos en determinar cómo un documento, D_j, que ha sido inspeccionado por el usuario, calcula el mensaje λ que manda a los términos T_i que lo indexan, y posteriormente, cómo éstos combinan los mensajes que reciben.

4.5.1.1. Cálculo de los mensajes λ.

Con objeto de determinar los valores concretos de los vectores λ, haremos la distinción entre documentos relevantes y no relevantes.

- *Documentos no relevantes.*

Como el documento ha sido considerado no relevante, mandará un mensaje $\lambda_{D_j}(T_i) = (1,0)$, indicando que el término no es útil para recuperar documentos relevantes, ya que podría introducir en las posiciones más altas documentos que no tienen que ver con la consulta. Se han probado otras técnicas alternativas, pero la que mejor se comporta es ésta que estamos describiendo, por lo que es la que tomaremos como vigente.

- *Documentos relevantes.*

  Vamos a considerar diferentes enfoques a la hora de calcular los vectores λ que los documentos relevantes mandarán a sus términos:

  d1: Consideraremos que todos los términos que contienen un documento que ha sido juzgado como relevante son también completamente relevantes, por lo que el documento les envía el mensaje $\lambda_{D_j}(T_i) = (0,1)$. Esta alternativa no es del todo válida, ya que a los términos de documentos relevantes se les asigna tanta importancia como a los términos de la consulta original, con el consiguiente peligro de cambiar el sentido inicial de la consulta.

  d2: En este caso, el vector λ enviado a los hijos es el siguiente:

  $$\lambda_{D_j}(T_i) = (p(d_j \mid \bar{t}_i), p(d_j|t_i))$$

  El cálculo de $p(d_j \mid t_i)$ se basará en el empleo de la función de probabilidad *fp10*, o lo que es lo mismo:

  $$p(d_j \mid t_i) = \sum_{k=1}^{m_j} w_{kj} \cdot p(t_k \mid t_i), \tag{4.11}$$

  donde w_{kj} es el peso del k-ésimo término en el j-ésimo documento, calculado según el producto $tf_{kj} \cdot idf_k^2$. m_j es el número de términos que indexan el documento D_j.

  Para estimar los distintos valores de $p(t_k \mid t_i)$ sería necesario realizar un gran número de propagaciones en la red documental, una por cada uno de los términos que indexan el conjunto de documentos inspeccionados, lo cual implicaría una tarea muy costosa en tiempo. Por tanto, para solucionar este problema se aproximan estas probabilidades haciendo la suposición de independencia entre términos, o lo que es lo mismo: $p(t_k \mid t_i) = p(t_k)$ si $i \neq k$ y $p(t_k \mid t_i) = 1$ si $i = k$.

  Teniendo en cuenta esta suposición, $p(d_j \mid t_i)$ se puede obtener de manera eficiente mediante la estimación de:

  $$p(d_j \mid t_i) = \sum_{k=1}^{m_j} w_{kj} \cdot p(t_k) + w_{ij}$$

  Así, $p(d_j \mid t_i) = p(d_j) + w_{ij} \cdot (1 - p(t_i))$. Análogamente se puede hacer el mismo razonamiento para calcular $p(d_j \mid \bar{t}_i)$, quedando finalmente $p(d_j \mid t_i) = p(d_j) - w_{ij} \cdot p(t_i)$

Observando los valores de los vectores λ calculados según estas expresiones anteriores, se detectó que eran muy cercanos a $(1,1)$, con lo que el efecto que se intenta producir no se está consiguiendo ya que prácticamente es equivalente a no instanciarlos. La razón para que ocurra esto es que $p(d_j \mid t_i)$ es bastante parecida a $p(d_j \mid \bar{t}_i)$. La solución a este problema pasa por realizar un cambio de escala, es decir, podríamos estimar $\lambda_{D_j}(T_i)$ como $\lambda_{D_j}(T_i)^{\delta}$, siendo $\delta > 1$. A esta variante de $d2$ pasaremos a denominarla $d2^{\delta}$.

d3: Este tercer enfoque que presentamos tiene en cuenta la influencia de la consulta original, Q, en los cálculos del vector λ, rompiendo así la primera restricción impuesta en el método anterior. Se trata, entonces, de estimar los mensajes como sigue a continuación:

$$\lambda_{D_j}(T_i) = (p(d_j \mid \bar{t}_i, Q), p(d_j \mid t_i, Q))$$

considerando cómo la probabilidad a posteriori de relevancia de un documento se ve afectada por medio de la adición de un nuevo término a la consulta. Utilizando la función de probabilidad *fp10*, la probabilidad $p(d_j \mid t_i, Q)$ se estima de la siguiente forma:

$$p(d_j \mid t_i, Q) = \sum_{k=1}^{m_j} w_{kj} \cdot p(t_k \mid t_i, Q) = \sum_{k=1, k\neq i}^{m_j} w_{kj} \cdot p(t_k \mid t_i, Q) + w_{ij}$$

El siguiente objetivo será estimar $p(t_k \mid t_i, Q)$ para cada término T_k que pertenezca a D_j. Estos cálculos requieren efectuar una propagación en la red documental considerando tanto la consulta como el nuevo término $T_i = t_i$ como evidencias, aunque de nuevo descartamos esta posibilidad por ser muy costosa en términos de tiempo. Así las cosas, intentaremos encontrar una aproximación con la información que poseemos. Consideramos que dos términos T_i y T_k, que indexan el documento D_j, y que han sido juzgados como relevantes, están correlados positivamente, aunque de una manera muy débil. Además, suponemos que el hecho de añadir T_i a la consulta original incrementará la evidencia en la relevancia de D_j. Estas dos suposiciones pueden ser combinadas utilizando para ello una puerta *OR*:

$$p(t_k \mid t_i, Q) = p(t_k \mid t_i, \bar{Q}) + (1 - p(t_k \mid t_i, \bar{Q}))p(t_k \mid \bar{t}_i, Q)$$

Por un lado, como $\bar{t}_i$ es el estado más probable que puede tomar T_i, es muy razonable suponer que no añade más información a la suministrada por Q, por lo que se puede hacer la aproximación: $p(t_k \mid \bar{t}_i, Q) \sim p(t_k \mid Q)$. Además, también se puede aproximar $p(t_k \mid t_i, \bar{Q}) \sim \varepsilon$, siendo ε un valor muy pequeño. La razón es que $p(t_k \mid t_i, \bar{Q})$ debería ser muy baja debido a que cuando $\bar{Q}$ se instancia, la probabilidad a posteriori de t_k debería ser muy pequeña y la influencia de instanciar t_i no contribuiría a incrementarla de forma notable. Teniendo en cuenta estas suposiciones y aproximaciones, la nueva expresión para estimar $p(t_k \mid t_i, Q)$ es:

$$p(t_k \mid t_i, Q) = \varepsilon + (1-\varepsilon)p(t_k \mid Q)) \tag{4.12}$$

Finalmente, tendríamos que:

$$p(d_j \mid t_i, Q) = \varepsilon \sum_{k=1}^{m_j} w_{kj} + (1-\varepsilon)\left(\sum_{k=1}^{m_j} w_{kj} \cdot p(t_k \mid Q) + w_{ij}(1 - p(t_i \mid Q))\right)$$

De manera análoga, $p(d_j \mid \bar{t}_i, Q)$ se calcula utilizando la expresión *fp10* y asumiendo que $p(t_k \mid \bar{t}_i, Q) \sim p(t_k \mid Q)$. Por tanto,

$$p(d_j \mid \bar{t}_i, Q) = \sum_{k=1}^{m_j} w_{kj} \cdot p(t_k \mid Q) - w_{ij} \cdot p(t_i \mid Q) \tag{4.13}$$

d4: En este método, la consulta Q juega el papel de un documento ficticio que ha sido observado como relevante. Por tanto, en el conjunto de documentos inspeccionados por el usuario figuran el documento D_j y la consulta Q, y su vector de verosimilitudes pasa a ser:

$$\lambda_{D_j}(T_i) = (p(d_j, Q \mid \bar{t}_i), p(d_j, Q \mid t_i))$$

Cabe destacar que este vector es equivalente, salvo una constante de normalización, a

$$\lambda_{D_j}(T_i) = \left(\frac{p(\bar{t}_i \mid d_j, Q)}{p(\bar{t}_i)}, \frac{p(t_i \mid d_j, Q)}{p(t_i)}\right),$$

que mide cómo cambia, considerando el documento y la consulta como relevantes, la creencia en la relevancia del término con respecto a la probabilidad a priori. Los numeradores de los cocientes anteriores se obtienen según:

$$p(t_i \mid d_j, Q) = \frac{p(d_j \mid t_i, Q) \cdot p(t_i, Q)}{p(d_j, Q)} = \frac{p(d_j \mid t_i, Q) \cdot p(t_i \mid Q)}{p(d_j \mid Q)}$$

Y $p(\bar{t}_i \mid d_j, Q))$ de manera análoga. Los valores $p(t_i \mid Q)$ y $p(d_j \mid Q)$ se calculan a partir de la consulta original, por lo que sólo se necesita determinar el valor de $p(d_j \mid t_i, Q)$, para lo cual utilizaremos dos métodos diferentes:

d4.1 Este método utiliza el mismo enfoque que *d3*.

d4.2 Teniendo en consideración la función *fp10*,

$$p(d_j \mid t_i, Q) = \sum_{k=1}^{m_j} w_{kj} \cdot p(t_k \mid t_i, Q),$$

asumiendo que $p(t_k \mid t_i, Q) = p(t_k \mid Q) + \beta$. De esta forma modelamos el hecho de que T_i y T_k sean casi independientes dado Q.

4.5.1.2. Combinación de los mensajes λ obtenidos por cada término.

Cuando varios documentos $D_1, \ldots, D_s$, tienen un cierto término T_i en común, el nodo asociado a dicho término recibirá un mensaje λ por cada uno de los documentos observados que lo contengan $\lambda_{D_j}(T_i), j = 1, \ldots, s$. Este conjunto de vectores λ debe ser combinado para conseguir una única medida $\lambda(T_i)$ que refleje la creencia en la relevancia o irrelevancia del término, combinación que se puede llevar a cabo de dos maneras alternativas:

- Una primera, donde no existe información extra de ningún tipo, con lo que la combinación se hace multiplicando todos los vectores λ que recibe un término en particular, combinación a la que llamaremos *simple*:

$$\lambda(T_i) = \prod_{j=1}^{s} \lambda_{D_j}(T_i). \tag{4.14}$$

- Y una segunda forma, la cual tiene en cuenta la calidad recuperadora de la consulta, medida como el número de documentos relevantes recuperados. La idea básica es que cuando el número de documentos relevantes recuperados en la consulta inicial es alto, ésta está haciendo un buen trabajo. Por tanto, parecería lógico que la consulta original se modificara poco, para lo cual los términos que se le añadieran no deberían tener un impacto muy fuerte en la consulta. Por otro lado, si la consulta original no recupera muchos documentos relevantes, la acción a realizar sería la contraria: añadir términos que tengan peso fuerte en esa consulta, aumentando la creencia en la relevancia de los términos que indexan esos documentos. La manera de poner en práctica esta idea pasa por utilizar una *combinación convexa*, donde n_r es el número de documentos relevantes recuperados y $\lambda(T_i)$ es el vector calculado utilizando (4.14):

$$\lambda'(T_i) = \alpha + (1-\alpha) \cdot \lambda(T_i), \text{ con } \alpha = \frac{n_r}{|J|} \tag{4.15}$$

La acción que realice cada nodo término cuando reciba los vectores de sus hijos dependerá del tipo de término que sea:

- Los términos de expansión positivos y negativos podrán aplicar directamente las ecuaciones (4.14) ó (4.15). Nótese que para los términos de expansión negativos, teniendo en cuenta que los documentos no relevantes les envían un vector $(1,0)$, el resultado es que el vector final utilizado por los términos es $(1,0)$.

- Los términos de expansión neutros (aparecen en documentos relevantes e irrelevantes al mismo tiempo), y debido a que el(los) documento(s) no relevante(s) donde aparecen le(s) manda(n) el vector $\lambda_D(T) = (1,0)$, es decir, se instancia el término a no relevante, pueden utilizar dos alternativas:

1. Descartar los mensajes, o lo que es lo mismo, no instanciar los términos, lo que implica el uso de un mensaje $\lambda(T_i) = (1,1)$.

2. Instanciarlos de acuerdo con sus distribuciones en los documentos relevantes o no relevantes de la ordenación; descartar todos los mensajes de documentos cuando el número de documentos no relevantes en los que aparezca el término $n_{\bar{r}t}$ sea igual al número de relevantes, n_{rt}, es decir, si $n_{rt} = n_{\bar{r}t}$ entonces $\lambda(T_i) = (1,1)$; descartar los mensajes de documentos relevantes cuando $n_{\bar{r}t} > n_{rt}$, en cuyo caso $\lambda(T_i) = (1,0)$; y por último, descartar los mensajes de los no relevantes cuando $n_{\bar{r}t} < n_{rt}$ y calcular $\lambda(T_i)$ mediante las combinaciones (4.14) o (4.15). Esta segunda opción, en experimentos preliminares produjo pésimos resultados, por lo cual fue descartada.

4.5.1.3. Experimentación sobre expansión de la consulta.

Comenzando con la exposición de los resultados conseguidos en la experimentación, y antes de pasar a comentarlos, indicar que el parámetro δ en la técnica $d2^{\delta}$ ha sido asignado a 5, ya que tras una experimentación previa es el valor con el que se alcanza un mayor rendimiento de los que se han probado; el parámetro ε utilizado para aproximar $P(t_k \mid t_i, \bar{Q})$ en *d3* se establece en 0,0075; y el valor asignado a β en el método *d4.2* será $1/|J|$, con $|J|$ el número de documentos recuperados. Todos estos valores se han obtenido como consecuencia de una fase previa de experimentación.

En la tabla 4.9 se muestran los resultados conseguidos con las diferentes alternativas diseñadas para hacer expansión de consultas en el método de realimentación de relevancia, basado en la instanciación de documentos con la colección CACM (los términos tc se instancian todos a $(0,1)$, los te- a $(1,0)$ y los te= a $(1,1)$). Tras la observación de dichos datos se pueden diferenciar tres comportamientos claramente definidos: el de la técnica $d1$, que demuestra que el considerar a todos los términos como completamente relevantes es una estrategia muy mala; el de $d2$ y su modificación $d2^{\delta}$, en este caso particularizado con $\delta = 5$, en el que se puede ver cómo el rendimiento cuando se emplea sin combinación convexa es mayor que cuando se utiliza ésta, notando además cómo al hacer el cambio de escala en el vector el rendimiento aumenta; y por último, el del grupo formado por $d3$ y las dos variantes de $d4$, con un comportamiento inverso al anterior grupo: funcionan considerablemente mejor cuando se aplica la combinación convexa, obteniendo unos porcentajes de mejora muy elevados con respecto a no usarla.

Si comparamos estos resultados con los obtenidos con el método basado en términos, $d2$ y $d2^5$ tienen un rendimiento menor que cualquiera de las técnicas $t2$, $t3$ y $t4$; al contrario que ocurre con $d3$ y las dos variaciones de $d4$, las cuales, utilizando la combinación convexa, superan a todas las técnicas salvo a la $t4$, que se mantiene como la mejor de todas en la fase de la expansión de consultas.

Exp.	te+	% de cambio combinación simple	% de cambio combinación convexa
1	$d1$	$-59{,}73$	
2	$d2$	33,27	30,87
3	$d2^5$	37,32	33,72
4	$d3$	$-40{,}19$	51,12
5	$d4{,}1$	$-40{,}18$	51,37
6	$d4{,}2$	$-46{,}18$	51,73

Cuadro 4.9.: Experimentos sobre expansión de consultas con el método basado en documentos con la colección CACM.

4.5.2. Repesado y expansión de la consulta.

El repesado de los términos de la consulta original se ha hecho instanciando los términos de la consulta positivos a $(0,1)$, al igual que los neutros, y los negativos se hace mediante la técnica $rc1$. En lo que se refiere a la expansión, los negativos se ponen a $(1,0)$, los neutros no se tienen en cuenta y por último, los positivos se instanciarán dependiendo de la técnica que se aplique. La asignación de valores a los parámetros permanece constante con respecto a la experimentación de la expansión.

Los resultados de esta experimentación se pueden observar en la tabla 4.10. En ella se aprecia cómo la combinación del repesado y de la expansión hace que se mejore ostensiblemente la recuperación con todas las técnicas de expansión. Destacar el aumento del porcentaje de cambio que se produce cuando la combinación de los vectores λ se hace simplemente multiplicándolos (aproximadamente un 30 %). También se da ese aumento, aunque más moderado (normalmente sobre un 20 %) cuando se usa la combinación convexa. En cuanto a las técnicas $d2$ y $d2^5$, la primera empeora casi un punto porcentual al aplicar la combinación convexa, y la segunda tiene la conducta contraria, mejorando algo más de un punto. En estas técnicas sigue sin apreciarse un cambio de rendimiento significativo entre las dos formas de combinar los mensajes. Por el contrario, donde sigue apreciándose esta diferencia es en los métodos $d3$ y $d4$ con sus dos variaciones, donde la aplicación de una combinación simple ofrece resultados negativos, aunque bastante más cercanos al cero que en los experimentos sin repesado. El mejor porcentaje de cambio se alcanza con $d4{,}1$, seguido a un punto de $d3$ y a dos y medio, $d4{,}2$, diferencias que podemos calificar como poco significativas.

Si comparamos estos resultados con los obtenidos con el método de realimentación basado en los términos, $d2$ y su variante son las peores técnicas, al contrario que $d3$ y $d4$ con sus dos variantes, que se comportan de forma muy parecida. La mejor de todas las técnicas sigue siendo $t4$.

Exp.	te+	% de cambio combinación simple	% de cambio combinación convexa
1	$d2$	54,85	53,68
2	$d2^5$	53,86	55,31
3	$d3$	−11,60	69,46
4	$d4,1$	−11,66	70,58
5	$d4,2$	−48,74	67,93

Cuadro 4.10.: Experimentos sobre repesado de la consulta original más expansión de consultas con el método basado en documentos con la colección CACM.

4.5.3. Selección de términos.

La siguiente fase de experimentación tiene por objeto comprobar cómo afecta el número de términos seleccionados para ser añadidos a la consulta original en la calidad recuperadora. Para ello, y al igual que se hizo en el método de realimentación basado en términos, se ha probado introduciendo 10, 20, 30 y todos los términos de expansión positivos (se hace repesado y expansión simultáneamente, tal y como se ha explicado en el apartado anterior, al mismo tiempo que utilizando la combinación convexa). En este caso, esta experimentación se ha realizado aplicando la técnica $d3$.

Para llevar a cabo estos experimentos, una vez calculado el vector λ final de cada término de expansión positivo, éstos se ordenan por el primer componente del vector (es un cociente de probabilidades, y por tanto, cuanto menor sea este valor más relevante será el término) y se selecciona el número de términos correspondiente.

De los porcentajes de cambio existentes en la tabla 4.11, se concluye que éstos van aumentando conforme se incrementa el número de términos que se añaden a la consulta, de tal forma, que el valor máximo, con nueve puntos porcentuales de diferencia con respecto al alcanzado añadiendo quince, se consigue cuando se suman todos los términos positivos a la consulta original.

En este caso, se produce un incremento del porcentaje de cambio con respecto al crecimiento del número de términos, al contrario que ocurría con el experimento análogo en el método de realimentación basado en los términos, donde se producía una disminución, aunque en ambos métodos se cumple que añadir todos los términos de expansión positivos ofrece una mayor efectividad, resultados contrarios a los alcanzados por Harman [Har92], donde demuestra empíricamente que es mejor añadir pocos términos a la consulta original.

4.5.4. Experimentación con las colecciones ADI, CISI, CRANFIELD y MEDLARS.

La tabla 4.12 contiene los porcentajes de cambio conseguidos al aplicar las diferentes técnicas pertenecientes al método de realimentación basado en los documentos sobre las colecciones

Exp.	Número de términos seleccionados	% de cambio
1	10	58,55
2	20	59,62
3	30	60,52
4	Todos	69,46

Cuadro 4.11.: Resultados de la selección de términos con la colección CACM y la técnica de expansión *d3*.

ADI, CISI, CRANFIELD y MEDLARS. En estos experimentos se han probado tanto la combinación simple (columnas etiquetadas como *% C.S.*) como la combinación convexa (*% C.C.*) en las cuatro colecciones.

La principal conclusión que se alcanza a la luz de estos porcentajes de cambio a través de las cuatro colecciones, más los ya conseguidos con la CACM, es que las tres técnicas del método que estamos tratando son bastante dependientes de la colección que se utilice como banco de pruebas.

Si nos centramos en cada una de las colecciones de manera individual, vemos que en ADI la técnica $d2^5$ alcanza el máximo porcentaje de cambio cuando se utiliza una combinación simple de los mensajes λ, y a la vez el valor más bajo de todos los experimentos al emplear la combinación convexa. Además, los mejores valores porcentuales se cruzan en $d2$ y en su variante: en la primera es mejor usar la combinación convexa, al contrario que en la segunda técnica. En cuanto a $d3$ y a las dos variantes de $d4$, en la primera y en $d4{,}1$ es preferible utilizar una combinación simple, aunque la diferencia no es muy grande con respecto a utilizar la convexa. Esta diferencia se mantiene también relativamente baja en $d4{,}2$, aunque es justo la contraria: mejor con combinación convexa, y además, el porcentaje de cambio con ambas combinaciones sube del orden de un veinte por ciento con respecto a las anteriores.

El comportamiento de CRANFIELD es más homogéneo con ambas combinaciones, ya que, salvo en la técnica $d2^5$, se mantiene por encima la combinación convexa con respecto a la simple (en $d2$ son prácticamente iguales). $d2$ también es la que se comporta peor, y $d4{,}2$ es la que consigue un rendimiento más alto.

CISI, con $d2$ y su variación, para las dos combinaciones, se comporta de manera muy parecida. Sin embargo, cuando nos fijamos en las técnicas $d3$ y las dos $d4$, se observa una clara mejoría, al igual que ocurre con la CACM, en favor del uso de la combinación convexa. Se observa cómo el rendimiento de la combinación simple baja considerablemente cuando se usan estas últimas técnicas en comparación con la familia de $d2$. Por último, comentar cómo los porcentajes de cambio se han visto decrementados en magnitud (aproximadamente a la mitad) con respecto a las dos colecciones ya comentadas anteriormente.

Para finalizar, MEDLARS es la que peores resultados ofrece. La razón fundamental puede ser que la red documental que hemos utilizado para propagar y la combinación de parámetros trabajan bastante bien, recuperando un número elevado de documentos relevantes en la primera

		ADI		CRANFIELD		CISI		MEDLARS	
Exp.	$te+$	% C.S.	% C.C.	% C.S.	% C.C.	% C.S.	% C.C.	% S.C.	% C.C.
1	$d2$	67,75	70,73	67,55	67,54	44,17	44,15	4,31	4,52
2	$d2^5$	137,70	64,88	87,81	70,32	43,64	44,53	−36,84	5,51
3	$d3$	85,86	82,98	82,30	90,53	2,99	48,49	−66,30	9,97
4	$d4,1$	85,85	78,76	82,30	93,46	2,98	45,76	−66,29	−11,97
5	$d4,2$	102,07	105,77	91,37	101,87	−7,72	42,51	−73,25	−4,79

Cuadro 4.12.: Porcentajes de cambio para el resto de colecciones con el método de realimentación basado en documentos.

consulta, por lo que la segunda consulta tiene poco que hacer, ya que dispone de poca información con la que trabajar. Aparte de este detalle, de nuevo la combinación convexa está por encima de la simple, con diferencias bastante elevadas (salvo en $d2$, donde son prácticamente iguales). En esta colección las técnicas $d3$ y la familia de $d4$ están muy por debajo de los valores positivos.

En general, podemos concluir que el uso conjunto del repesado y la instanciación de los documentos (expansión) es una combinación bastante útil para nuestro modelo (como ya ha sido confirmado en otros modelos en [Har92, SB90]). Además, la aplicación de la combinación convexa parece aportar normalmente un mejor rendimiento, aunque para las técnicas $d3$ y $d4$, en la mayoría de las colecciones, es una ayuda necesaria, al contrario que ocurre con $d2$ y su variación, en la que ambas combinaciones ofrecen porcentajes de cambio muy parecidos. Los resultados con este método están en la línea de los conseguidos por Harman, Salton y Buckley en las referencias anteriores, salvando las distancias producidas por los entornos experimentales diferentes.

Comparando este método de realimentación con el basado en términos, podemos decir que, para las colecciones CRANFIELD y MEDLARS, parece que este segundo tiene un comportamiento mejor. Esta situación se invierte cuando se usa la combinación convexa con el método basado en documentos cuando consideramos la colección CISI, donde su rendimiento se incrementa con respecto al basado en términos. Con ADI, el decantarse por un método es más complicado a la luz de los resultados ofrecidos en las tablas 4.8 y 4.12.

5. Extensión del modelo de red bayesiana documental: redes de documentos.

El hecho de suponer que los documentos no están relacionados entre sí directamente, como ocurre en los modelos de red bayesiana documental desarrollados en el capítulo 3 es muy restrictivo, ya que en cualquier colección existen documentos que tratan del mismo tema y, por tanto, están estrechamente ligados. El problema radica en que, una vez que un documento se ha "reducido" a una representación mediante una lista de términos de indexación, resulta difícil establecer relaciones entre documentos si no es a través de los términos. Es por ello que la topología de las redes bayesianas utilizadas establecen enlaces entre nodos término y nodos documento. Así, ante una consulta, se puede dar el caso de que se recupere un documento que se considera relevante a ella, y que no se recupere otro documento que está íntimamente relacionado con el primero, y que por cualquier causa (por ejemplo, un método de indexación diferente, o el uso de términos de indexación distintos aunque semánticamente similares) el S.R.I. ha considerado que no tiene un grado de relevancia suficiente para ser también entregado al usuario.

Por tanto, puede ser importante para el rendimiento del S.R.I. el incluir las relaciones directas entre documentos, tratando de conseguir así una mejoría en las curvas de exhaustividad - precisión.

En esta sección presentamos el método desarrollado para llevar a cabo esta tarea, el cual se basa en la construcción de una subred de documentos generada a partir de la puesta en práctica de un método de agrupamiento de documentos (*clustering*), utilizando para ello la propia red bayesiana.

5.1. Introducción al agrupamiento en recuperación de información.

Basado en la *hipótesis del agrupamiento* (en inglés, *clustering hypothesis*), la cual establece que los documentos que versen sobre el mismo tema tenderán a ser relevantes a la misma consulta [Rij79], el agrupamiento de documentos consiste en determinar conjuntos o clases de documentos en los que cada ítem de la colección pueda clasificarse según el grado de afinidad con respecto a los elementos que haya en cada grupo. El objetivo fundamental es el de mejorar la calidad de la recuperación. Así, una vez establecidos los grupos, la recuperación ganará en velocidad debido a que la consulta sólo se comparará con el representante de cada grupo, conocido como *centroide*, es decir, un documento ficticio que representa el contenido de todos los documentos integrados en ese grupo. Además, el rendimiento de la recuperación se ve incrementado, debido a que pueden recuperarse documentos que son relevantes totalmente a la consulta, pero que no contienen en común ningún término.

Existe una gran cantidad de algoritmos para llevar a cabo esta tarea [Rij79, SM83, Ras92], pero todos se fundamentan, de una u otra forma, en la existencia de una función de similitud que mide el grado de asociación entre dos documentos (ejemplos de estas funciones pueden ser la medida del coseno [SM83] o la medida de Kullback-Leibler [KL51]). Para cada par de documentos se calcula el grado de similitud, y posteriormente, dado un umbral, se introducen en el mismo grupo aquellos documentos que tengan una similitud mayor o igual que dicho umbral, estableciendo así de forma implícita las relaciones entre documentos.

En lo que se refiere a cómo mediante redes bayesianas se han representado las relaciones entre los documentos de la colección, la extensión de la red de documentos que hacen Croft y Turtle [Tur90] pasa por agrupar los documentos en clases. Cada clase se representará en la red por un nuevo nodo, al que apuntarán todos los nodos documentos que están contenidos en la clase. Los nuevos nodos de clases que se incorporan a la red representan a los documentos centroides de sus propias clases, enlazándose con los términos que contienen como cualquier otro documento. Además, ponen en práctica dos técnicas para relacionar documentos: la utilización de las *citas entre documentos* y los *enlaces a los vecinos más próximos*. La primera de ellas parte del hecho de que se dispone de la lista de documentos que se citan en cada uno de los documentos de la colección. Se puede enlazar, por tanto, un documento D_i con todos aquellos que son citados en él. La segunda técnica calcula la similitud de cada documento con el resto y selecciona los que son más parecidos a él, estableciéndose un enlace entre los nodos que representan a dichos documentos.

No exactamente en el campo del agrupamiento, pero sí en el cercano de la clasificación documental (dadas una categorías temáticas fijas, la asignación de los documentos a alguna de éstas), Sahami [Sah98] ha desarrollado un algoritmo de clasificación basado en el aprendizaje de una red bayesiana, consiguiendo porcentajes de acierto bastante alto.

5.2. Construcción de la subred aumentada de documentos.

El objetivo que nos planteamos en esta sección es la consecución de un conjunto de relaciones entre documentos con el objeto de enriquecer y aumentar la capacidad de recuperación del modelo. Este fin se alcanzará utilizando un razonamiento análogo al que se usa en la red bayesiana documental para obtener una lista ordenada de los relevantes a una consulta.

Dada una consulta, que en términos de nuestro modelo se representa como un conjunto de evidencias, con la propagación de evidencias se consigue calcular el grado de relevancia de cada documento con respecto a la consulta efectuada al S.R.I., es decir, la probabilidad de que cada documento sea relevante dado que las variables término de la consulta son también relevantes. Estos valores se pueden ver como el grado de cercanía de cada documento con respecto a la consulta.

Entre un documento y una consulta no existen diferencias de representación: ambos son vectores que contienen los términos por los que han sido indexados los textos que representan. Por esta razón, podemos pensar en instanciar, en lugar de la consulta, los términos de cada uno de los documentos de la colección, consiguiendo así, para cada uno de ellos, una lista ordenada de documentos según su probabilidad de relevancia con respecto al documento instanciado en cada momento. Para un documento concreto que actúe como evidencia, los documentos que estén en posiciones más altas de la ordenación generada, es decir, los que tienen una mayor probabilidad condicionada, serán aquellos que estén más relacionados con él. De esta manera, se puede conseguir para cada documento el conjunto de documentos con los que tiene más similitudes, o visto desde otra perspectiva, con los que está más estrechamente relacionado.

Estas relaciones entre documentos se representarán en la subred de documentos como arcos que se originan en el nodo documento que actúa como instancia y apuntan a los nodos documento que tienen un mayor valor de probabilidad de relevancia con respecto a él. La nueva subred de documentos está formada, por tanto, por dos capas de documentos, conteniendo cada una de ellas todos los documentos de la colección. Las dos capas se unen únicamente con arcos que unen una capa a la otra.

En la figura 5.1 se presenta un ejemplo de la nueva subred de documentos. En él se aprecia un total de 8 documentos, de los cuales sólo cuatro actúan de padres del resto (sólo se han enlazado cuatro nodos documento con el resto por motivos de claridad).

La cuestión que inmediatamente surge a partir de esta forma de proceder es la manera de determinar el número de documentos con los que relacionar cada uno de los instanciados. Las soluciones serían dos:

- Tomar un umbral de probabilidad a posteriori, de tal forma que todos los documentos que tuvieran una probabilidad mayor que dicho umbral fueran incorporados a la subred de documentos como hijos del documento que se ha instanciado.
- Utilizar un número fijo de documentos con los que relacionarlos, es decir, los c más relevantes.

En nuestro caso, hemos optado por la segunda opción, aunque lo ideal sería poder establecer el umbral en la probabilidad o el límite en el número de documentos a partir de información relacionada con la colección con la que se está trabajando y, al mismo tiempo, contar con información sobre cada documento específicamente. Esta tarea quedará como una línea de investigación abierta que será tratada en el futuro.

El algoritmo 5.1 muestra el conjunto de pasos que hay que realizar para ampliar la red bayesiana documental con las relaciones entre documentos.

Como se puede apreciar en dicho algoritmo, hay tres partes claramente diferenciadas:

Algoritmo 5.1 Construcción de la subred de documentos.

entradas: red bayesiana documental (*RBD*) y número de documentos con los que relacionar (ndr) cada documento.
salidas: la red bayesiana documental ampliada con los documentos de la colección conectados entre sí.
constante: $CORTE \leftarrow 15$ {Número máximo de documentos con los que construir el fichero invertido.}
{Instanciación de cada documento y elaboración de la lista de documentos}
para cada documento $D_j \in RBD$ **hacer**
  Instanciar los términos $T_i, i = 1, \ldots, m_j$ pertenecientes al documento D_j a relevante.
  Efectuar la propagación de evidencias y obtener $p(d_k \mid d_j), \forall D_k \in RBD$.
  Generar una lista ordenada de documentos según dicha probabilidad y seleccionar los $CORTE$ más relevantes. Esta lista se denominará O_{D_j} y contendrá los documentos seleccionados y la probabilidad correspondiente $p(d_k \mid d_j)$ de cada uno de ellos.
fin para
Crear una nueva capa de nodos documento, *ND*, en *RBD*.
para cada documento $D_j \in ND$ **hacer**
  $FI_{D_J} \leftarrow \emptyset$
  para cada documento $D_i \in RBD$ **hacer**
    si $D_j \in O_{D_i}$ **entonces**
      $FI_{D_j} \leftarrow FI_{D_j} \cup (D_i, p(d_j \mid d_i))$
    fin si
  fin para
fin para
para cada documento $D_j \in ND$ **hacer**
  Ordenar FI_{D_j} decrecientemente según las probabilidades $p(d_j \mid d_i)$, con $D_i \in FI_{D_j}$ y seleccionar los ndr primeros documentos. El resto, eliminarlos.
  Conectar $D_i \in RBD$ con $D_j \in ND$ mediante un arco dirigido del primero al segundo, es decir, $D_i \rightarrow D_j$.
fin para
Devolver *RBD*.

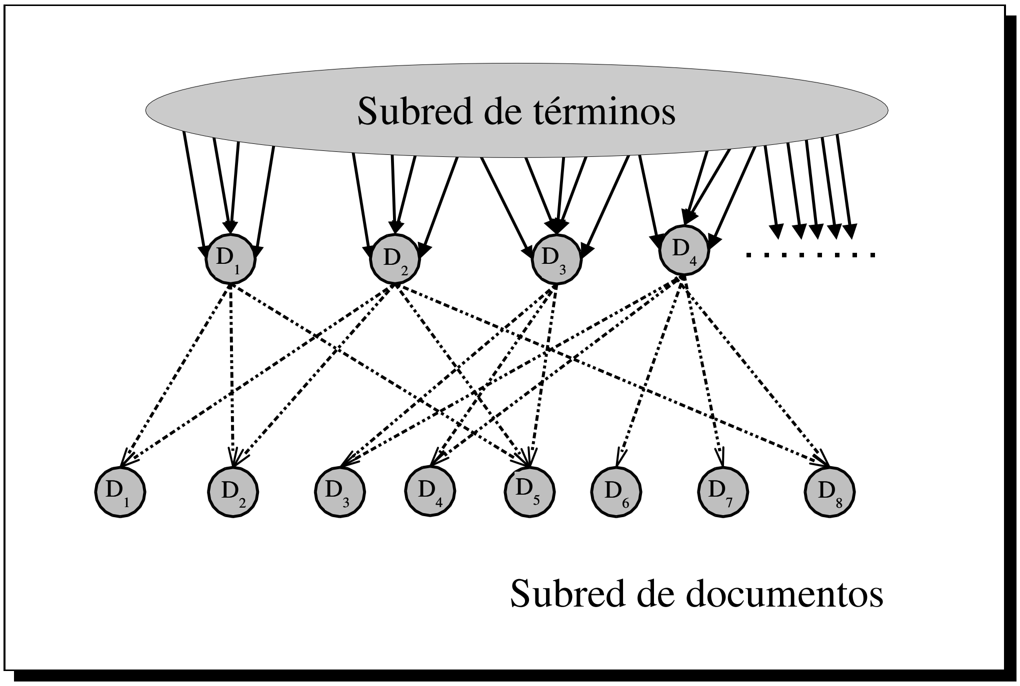


Figura 5.1.: Ejemplo de una subred de documentos con relaciones entre éstos.

- Generación de las listas ordenadas de documentos según su relevancia a cada documento D_j de la colección, O_{D_j} (pasos 4 a 8).

  Para ello, se instancian a relevante los términos de cada documento, se propaga y se ordenan los documentos de manera decreciente según la probabilidad a posteriori calculada. En estas ordenaciones no aparecen todos los documentos, sino un número limitado de éstos. En nuestro caso, y por analogía al número de documentos que devuelve el S.R.I. al usuario tras una consulta, es 15. Esta limitación sólo se hace con vistas a ahorrar tiempo de cálculo.

  Veamos el ejemplo correspondiente a la figura 5.1. La lista siguiente representa los documentos más relevantes con respecto a cada documento de la colección (en este ejemplo, la colección tiene ocho documentos).

$$
\begin{aligned}
O_{D_1} &= ((D_1, 1{,}0), (D_2, 0{,}8), (D_5, 0{,}7)) \\
O_{D_2} &= ((D_2, 1{,}0).(D_1, 0{,}9), (D_5, 0{,}8).(D_8, 0{,}6) \\
O_{D_3} &= ((D_3, 1{,}0), (D_4, 0{,}9), (D_5, 0{,}9)) \\
O_{D_4} &= ((D_4, 1{,}0), (D_3, 0{,}9), (D_6, 0{,}8), (D_7, 0{,}7), (D_8, 0{,}5)) \\
O_{D_5} &= ((D_5, 1{,}0), (D_6, 0{,}8), (D_4, 0{,}8), (D_3, 0{,}4)) \\
O_{D_6} &= ((D_6, 1{,}0), (D_7, 0{,}6), (D_3, 0{,}5), (D_8, 0{,}4) \\
O_{D_7} &= ((D_7, 1{,}0), (D_4, 0{,}7), (D_2, 0{,}7), (D_1, 0{,}1)) \\
O_{D_8} &= ((D_8, 1{,}0), (D_7, 0{,}8), (D_1, 0{,}8), (D_3, 0{,}8))
\end{aligned}
$$

  Al instanciar los términos de un documento, y posteriormente propagar, la probabilidad a posteriori del documento cuyos términos se han instanciado debe ser 1,0.

- Generación de un fichero invertido para todos los documentos (pasos 9 a 17).

  Para cada documento, D_j, se seleccionan los documentos D_i para los que D_j se ha considerado relevante. Continuando con el ejemplo, el fichero invertido completo sería el siguiente:

$$
\begin{aligned}
FI_{D_1} &= ((D_1, 1{,}0), (D_2, 0{,}9), (D_8, 0{,}8), (D_7, 0{,}1)) \\
FI_{D_2} &= ((D_2, 1{,}0), (D_1, 0{,}8), (D_7, 0{,}7)) \\
FI_{D_3} &= ((D_3, 1{,}0), (D_4, 0{,}9), (D_8, 0{,}8)), (D_6, 0{,}5), (D_5, 0{,}4) \\
FI_{D_4} &= ((D_4, 1{,}0), (D_3, 0{,}9), (D_5, 0{,}8), (D_7, 0{,}7)) \\
FI_{D_5} &= ((D_5, 1{,}0), (D_3, 0{,}9), (D_2, 0{,}8), (D_1, 0{,}7)) \\
FI_{D_6} &= ((D_6, 1{,}0), (D_4, 0{,}8), (D_5, 0{,}8)) \\
FI_{D_7} &= ((D_7, 1{,}0), (D_8, 0{,}8), (D_4, 0{,}7), (D_6, 0{,}6)) \\
FI_{D_8} &= ((D_8, 1{,}0), (D_2, 0{,}6), (D_4, 0{,}5), (D_6, 0{,}4))
\end{aligned}
$$

  Como se puede apreciar, el documento D_j siempre está relacionado consigo mismo, incluyéndose en la primera posición del fichero invertido.

- Selección de los documentos más relacionados con cada documento de la colección e inserción de los arcos correspondientes en la red bayesiana documental (pasos 18 a 21).

  Una vez creado el fichero invertido, se ordenan los documentos contenidos en cada una de las entradas del mismo según las probabilidades a posteriori. Seguidamente, se seleccionan los más relevantes (según el valor de la variable ndr) para cada documento (entrada). Éstos serán sus padres en la red bayesiana documental. Nótese que el valor ndr está imponiendo un límite al número de padres que pueda tener un documento, no respecto al número de hijos (que está limitado a un máximo de 15 en los experimentos que hemos realizado, únicamente para que el proceso de creación del fichero invertido sea más rápido).

  En el ejemplo, con el que estamos ilustrando esta sección, se ha fijado ndr a 3 y tras establecer el corte, el fichero invertido final (los documentos que pertenecen a cada conjunto FI_{D_j} serán los padres de D_j):

$$
\begin{aligned}
FI_{D_1} &= ((D_1, 1{,}0), (D_2, 0{,}9), (D_8, 0{,}8)) \\
FI_{D_2} &= ((D_2, 1{,}0), (D_1, 0{,}8), (D_7, 0{,}7)) \\
FI_{D_3} &= ((D_3, 1{,}0), (D_4, 0{,}9), (D_8, 0{,}8)) \\
FI_{D_4} &= ((D_4, 1{,}0), (D_3, 0{,}9), (D_5, 0{,}8)) \\
FI_{D_5} &= ((D_5, 1{,}0), (D_3, 0{,}9), (D_2, 0{,}8)) \\
FI_{D_6} &= ((D_6, 1{,}0), (D_4, 0{,}8), (D_5, 0{,}8)) \\
FI_{D_7} &= ((D_7, 1{,}0), (D_8, 0{,}8), (D_4, 0{,}7)) \\
FI_{D_8} &= ((D_8, 1{,}0), (D_2, 0{,}6), (D_4, 0{,}5))
\end{aligned}
$$

  Así, por ejemplo, el documento D_4 se ha encontrado, dentro del corte correspondiente en un total de tres documentos (el D_3 y D_5) además de él mismo.

Comentar, por último, dos peculiaridades importantes. Por un lado, la subred de documentos queda ahora compuesta por dos capas idénticas de documentos, en vez de una sóla como se confeccionó inicialmente. Este hecho da una gran versatilidad, ya que la segunda capa y los arcos que unen documentos de ésta con los de la primera se añaden de acuerdo con el hecho de si se desea recuperar con ella o no. Además, de esta manera se mantiene la topología de un G.D.A. y nos permitirá utilizar una técnica de propagación similar a la evaluación de las funciones de probabilidad en la red bayesiana documental original.

La segunda cuestión es que un documento de la segunda capa siempre tendrá como padre al documento homónimo de la primera capa. La probabilidad asociada será $p(d_j \mid d_j) = 1$. Esto se hace así con objeto de que la ordenación de documentos generada tras la propagación en la nueva subred de documentos mantenga, en cierta medida, la ordenación conseguida tras la propagación en la red bayesiana documental original. Así se evita que la ordenación se modifique de forma drástica.

5.3. Estimación de las distribuciones de probabilidad condicionadas en los nodos documento de la nueva capa.

Como a cualquier otro nodo de la red bayesiana documental, a los nodos documento de la nueva capa hay que estimarles la distribución de probabilidad condicionada dados sus padres. Siguiendo la filosofía de las funciones de probabilidad, y por las mismas razones que se pusieron en práctica en el capítulo 3, la que se ha diseñado para esta ocasión es la siguiente:

- **fpd1**:

$$p(d_j \mid \pi'(D_j)) = \frac{1}{\sum_{i=1}^{m'_j} p(d_j \mid d_i)} \sum_{D_i \in R_{\pi'(D_j)}} p(d_j \mid d_i), \tag{5.1}$$

siendo m'_j el número de padres del documento D_j (es decir, el número de documentos con que está relacionado) y $\pi'(D_j)$ una configuración para los padres documento de D_j. Esta expresión representa una suma de las probabilidades condicionadas de dicho documento dada cada documento que esté a relevante en la configuración, normalizada por la suma de las probabilidades de todos los documentos.

5.4. Recuperación de documentos con la nueva subred de documentos.

En esta sección vamos a centrarnos en la forma en que se van a actualizar las probabilidades a posteriori de los documentos dada la consulta a la luz de las relaciones existentes entre los documentos de la colección.

Para ello, el proceso de inferencia se lleva a cabo en dos partes:

- Instanciación de los términos de la consulta en la red bayesiana documental (la subred de documentos inicialmente estará formada por los documentos sin ninguna relación directa entre ellos) y cálculo de las probabilidades de relevancia de cada documento dada la consulta.
- Combinación de la probabilidad a posteriori obtenida en la fase anterior con la información disponible en la nueva subred de documentos, es decir, las probabilidades de relevancia de un documento dado otro, actualizando así esta primera probabilidad a la luz de las relaciones documentales generadas.

Se ha diseñado el siguiente método para llevar a cabo esta combinación de informaciones:

Una vez que se ha propagado, para cada documento se tiene la probabilidad de que el documento sea relevante dada la consulta. Por otro lado, de la fase de creación de la subred de documentos se dispone del fichero invertido de documentos, conteniendo las probabilidades de relevancia de cada documento dado que un segundo es también relevante. Con estos dos tipos de probabilidades podemos generar una única, a partir de la expresión (5.1), de la siguiente manera:

$$p(d_j \mid Q) = \frac{1}{\sum_{i=1}^{ndr} p(d_j \mid d_i)} \sum_{i=1}^{ndr} p(d_j \mid d_i) \cdot p(d_i \mid Q) \tag{5.2}$$

O lo que es lo mismo, la probabilidad de relevancia dada la consulta de cada documento D_i con el que D_j esté relacionado en el fichero invertido, se pondera por la probabilidad de relevancia de D_j dado que D_i es relevante. Además, como el propio documento D_j está relacionado consigo mismo, la probabilidad de relevancia del documento dada la consulta también interviene en la sumatoria.

Podemos garantizar que al utilizar *fpd1*, los resultados de la propagación serán exactos, esto es, se cumplen los requisitos del teorema 3.1, esto es, se puede expresar de la siguiente forma:

$$p(d_j \mid Q) = \sum_{i=1}^{ndr} wd_{ij} \cdot p(d_i \mid Q),$$

siendo wd_{ij} un peso del documento D_i con el que D_j está relacionado en el fichero invertido. Este peso toma la forma:

$$wd_{ij} = \frac{p(d_j \mid d_i)}{\sum_{i=1}^{ndr} p(d_j \mid d_i)}$$

EXP.	SMART	Mejor res.	$ndr = 5$	$ndr = 10$	$ndr = 15$
ADI					
REL. REC.	91	91	92	96	94
M. 11PTS	0.4709	0.4706	0.5023	**0.5095**	0.5035
%C. 11PTS		0.1	6.7	**8.2**	6.9
CACM					
REL. REC.	246	244	222	233	236
M. 11PTS	0.3768	0.4046	0.3602	0.3572	0.3547
%C. 11PTS		7.3	-4.5	-5.3	-5.9
CISI					
REL. REC.	343	369	363	370	376
M. 11PTS	0.2459	0.2642	0.2661	**0.2802**	0.2769
%C. 11PTS		7.4	8.2	**13.9**	12.6
CRANFIELD					
REL. REC.	824	836	543	576	574
M. 11PTS	0.4294	0.4434	0.2183	0.2318	0.2450
%C. 11PTS		3.2	-49.2	-46.1	-43.0
MEDLARS					
REL. REC.	260	287	266	279	269
M. 11PTS	0.5446	0.6443	0.5631	0.5799	0.5566
%C. 11PTS		18.3	3.3	6.4	2.2

Cuadro 5.1.: Evaluación de la extensión de la subred de documentos para todas las colecciones.

5.5. Experimentación con el nuevo modelo.

La experimentación que hemos realizado para determinar el rendimiento de las propuestas anteriores en las cinco colecciones con las que estamos trabajando ha sido la siguiente:

- El fichero invertido de documentos se ha generado con tres valores de corte ndr distintos: 5, 10 y 15.
- Para cada fichero invertido generado y la red que ofrece un mejor rendimiento en cada colección, según la experimentación llevada a cabo en el capítulo 3 (tabla 3.20), se ha aplicado la expresión *fpd1*.

En la tabla 5.1 se muestran los resultados resumidos que se han conseguido con cada colección para cada uno de los valores de corte, además de los pertenecientes al mejor modelo para cada colección (todos los porcentajes de cambio hacen referencia a SMART). Las tablas completas se pueden ver en el anexo C, así como las curvas Exhaustividad - Precisión..

Sólo dos (ADI y CISI) de las cinco colecciones mejoran los resultados obtenidos con la red bayesiana documental sin conectar los nodos documento. Además, los incrementos son bastante sustanciales estableciéndose como nuevos máximos de porcentaje de cambio. Estas dos colecciones, ADI y CISI, coinciden en la característica de que sus mejores resultados se obtienen con la red simple. También se aprecia que para CACM y MEDLARS tampoco hay mejoría (estos experimentos se han efectuado con redes aumentadas). Por último, CRANFIELD es la que peores resultados ofrece con diferencia (la única cuyas pruebas se realizaron con la red mixta).

Con respecto a SMART, son tres las colecciones donde la subred de documentos extendida siguen superándola: ADI, CISI y MEDLARS.

Por otro lado, ADI, CISI y MEDLARS tienen una conducta similar con respecto al valor utilizado para el corte: el máximo lo tienen en el valor 10. En las dos primeras el porcentaje de cambio de la media de la precisión para los once puntos de exhaustividad es mayor con un corte de quince documentos con respecto al de cinco. Se da la situación opuesta para MEDLARS.

Una posible explicación a este comportamiento (mejoría con la red simple y empeoramiento con el resto) puede ser la siguiente: tanto en la red aumentada como mixta existen relaciones explícitas entre términos en la subred de términos. Dichas relaciones se reflejan, en el momento de la propagación en la subred de términos, en la probabilidad a posteriori de cada término dada la consulta. Un proceso análogo es el que se hace cuando se desea determinar qué documentos son los más relevantes para uno concreto: se instancian los términos del documento y se propaga, calculando la probabilidad a posteriori de cada documento dado el documento en cuestión. Cuando se combinan estos dos tipos de redes con la subred extendida de documentos, el proceso de propagación se ha efectuado dos veces en la red y podría darse el caso de que la información resumida en las probabilidades a posteriori correspondientes fuera de alguna forma "semejante", produciéndose una duplicidad que parece que no es buena.

Cuando se recuperase con la red simple y la extensión de la subred de documentos, no se produciría esta redundancia, pues en la primera no existen relaciones directas entre términos.

La función de probabilidad *fpd1* podría ser objeto de varias modificaciones para tratar de mejorar su rendimiento. Una de éstas, y estableciendo la analogía con las funciones de probabilidad que combinan las probabilidades a posteriori de los términos que contienen, podría ser la introducción de las diferencias de probabilidades del documento. En la experimentación realizada con los diferentes tipos de red bayesiana documental expuesta en el capítulo 3, se ha visto que es una técnica bastante útil, pues aumenta notoriamente la calidad del S.R.I. En el caso que nos ocupa, en vez de utilizar la probabilidad a posteriori del documento dada la consulta en la expresión (5.2), se podría utilizar $p(d_i \mid Q) - p(d_i)$. La probabilidad a priori se obtendría tras un proceso de propagación sin evidencias en la red bayesiana documental original (sin extensión de la subred de documentos).

6. Conclusiones y trabajos futuros.

6.1. Conclusiones.

En esta memoria hemos presentado un modelo de recuperación de información totalmente basado en redes bayesianas. Por un lado, se ha especificado completamente, exponiendo de manera detallada tanto la topología de las redes que lo sustentan (la información cualitativa) como la forma en que se estiman las diferentes distribuciones de probabilidad almacenadas (información cuantitativa). El siguiente paso que se ha dado ha sido el establecer un mecanismo eficiente que nos permita realizar la inferencia cuando el usuario efectúa una consulta al S.R.I., y poder asociar así a cada documento un grado de relevancia con respecto a esa consulta.

La Recuperación de Información estudia un problema enormemente complejo, no sólo por la incertidumbre intrínseca asociada a muchos aspectos del problema (relacionados básicamente con la representación de documentos y consultas), sino también por el propio tamaño del problema, en cuanto al número de elementos que lo componen (documentos y términos). Esto último ha supuesto un reto importante desde el punto de vista de la armonización, por un lado del uso de mecanismos de representación e inferencia sólidamente fundamentados, y por otro, de la necesidad de poner en práctica esos métodos en condiciones aceptables en cuanto al tiempo de respuesta. En ese sentido, el desarrollo de esta memoria ha supuesto una lucha constante contra la complejidad.

Por tanto, hemos tenido que desarrollar diversas técnicas que, sin perder el fundamento teórico que aportan los modelos gráficos probabilísticos, fueran capaces de enfrentarse a la complejidad de los problemas considerados:

- Así, para el problema del aprendizaje de la subred de términos, tuvimos que restringir nuestro algoritmo a un tipo de redes más simplificadas (los poliárboles), con objeto de obtener tiempos de aprendizaje de la red y, sobre todo, tiempos de inferencia con ella aceptables. Pero incluso restringiéndonos a este tipo de redes, hubo que diseñar un algoritmo específico, que combinase metodologías útiles de otros algoritmos existentes, e incorporase características particulares[1]. Esto ha sido necesario porque, en pruebas preliminares con otros algoritmos de aprendizaje de poliárboles, fue imposible obtener resultados[2].

[1]Para limitar, por ejemplo, el número de padres de los nodos mediante tests de independencia.

[2]A título de ejemplo, la ejecución de uno de estos algoritmos sobre la colección más sencilla de todas (ADI),

- Para la estimación de la información cuantitativa (distribuciones de probabilidad condicionadas), nos encontramos con el problema del enorme número (con respecto a lo que es habitual en otros modelos de redes bayesianas empleados en otros ámbitos) de nodos padres que los nodos documento poseen en la red bayesiana documental. Esto hacía imposible cualquier intento de estimación, almacenamiento y utilización de estas distribuciones. Por tanto, hubo que diseñar también técnicas específicas, las denominadas funciones de probabilidad. Estos métodos, similares a lo que en la literatura sobre redes bayesianas se denominan modelos de interacción canónica o modelos de independencia causal, permiten en primera instancia calcular y almacenar (de forma implícita) todas las probabilidades condicionadas, en función de un número muy limitado de parámetros o probabilidades más elementales. Además de proponer un nuevo modelo de interacción canónica genérico, todos los modelos utilizados se han adaptado a las características propias de la disciplina de la recuperación de información, empleando para ello medidas habituales en este campo [3].

- Para a continuación poder emplear nuestros modelos era necesario utilizar mecanismos de inferencia (cálculo de probabilidades a posteriori). Los métodos habituales de inferencia en redes bayesianas se demostraron completamente insuficientes para tratar con la envergadura de nuestros modelos. Ha sido necesario, por tanto, desarrollar una técnica específica de inferencia, que saca partido de la topología concreta de nuestra red bayesiana documental y de nuestras funciones de probabilidad, para conseguir realizar inferencia exacta en una red globalmente compleja, dividiendo el proceso en una primera fase de inferencia exacta en una red sencilla (la subred de términos), y a continuación otra etapa de evaluación de las funciones de probabilidad.

En el campo de la Recuperación de Información, como en muchos otros, los modelos formales y las propuestas teóricas carecen de mucho valor si no se acompaña de pruebas empíricas que los avalen. En este sentido, el trabajo experimental desarrollado en esta memoria ha pretendido ser a la vez amplio y riguroso: para todas las propuestas teóricas planteadas se han realizado las correspondientes pruebas experimentales. Además, en la mayoría de los casos, los experimentos se han llevado a cabo con cinco de las colecciones estándar de prueba más comúnmente utilizadas, con características muy diferentes, tratando de validar nuestras propuestas en diferentes situaciones, evitando sesgos y sobreajustes que hubieran podido producirse experimentando, por ejemplo, con una única colección. También se ha tratado de huir de triunfalismos, los resultados obtenidos se exponen en toda su crudeza, tanto si son buenos como si no lo son. En muchas ocasiones los resultados experimentales negativos nos han forzado a profundizar más para tratar de mejorar nuestras propuestas iniciales.

El modelo presentado ofrece una gran versatilidad en cuanto a su uso, pues puede ser utilizado de tres maneras diferentes:

tuvo que ser abortada, y se pudo estimar que hubiese necesitado cuatro años para aprender completamente la estructura.

[3]Fórmulas basadas en los valores tf o idf, similares a la típicas fórmulas del coseno o el coeficiente de Jaccard.

- Como un subsistema de expansión de consultas para cualquier S.R.I. externo. En esta modalidad, se explota la gran capacidad que posee el modelo bayesiano sobre el que está basado para inferir nueva información a partir de las relaciones entre los términos de la colección. Considerando esta nueva información, le suministra al sistema un conjunto de términos adicionales, relacionados con los de la consulta, con objeto de ampliar la expresividad de la misma.

- Como un S.R.I. autónomo. En este caso, el modelo desarrollado actúa como un modelo de recuperación de información propiamente dicho, ya que es él quien, de forma autónoma, determina el grado de semejanza de los documentos con respecto a la consulta.

- Como un mecanismo de mejora de la consulta. Se le ha dotado de una metodología para realizar realimentación de relevancia, permitiendo así, en colaboración directa con el usuario, que se refinen las consultas de manera que éste finalice lo más satisfecho posible.

Las conclusiones generales que podemos sacar del trabajo realizado y de los resultados obtenidos son las siguientes:

- A pesar de que el marco formal que fundamenta las redes bayesianas ha sido desarrollado de manera general, éstas presentan la gran ventaja de que pueden ser utilizadas convenientemente en problemas muy específicos, consiguiendo una alta calidad en las soluciones obtenidas. La adaptación a dichos problemas se puede lograr mediante la modificación, sin mucho esfuerzo, de algunas técnicas básicas. Este es el caso de la Recuperación de Información, donde, a pesar de que ya ha sido concluido por otros autores, corroboramos la total idoneidad del formalismo basado en redes bayesianas para ser aplicado con éxito a los problemas existentes en esta disciplina.

- El modelo desarrollado presenta una gran versatilidad porque puede ser utilizado en diferentes ámbitos de la R.I., como ya hemos dicho anteriormente, pero con la característica adicional que permite que cada uno de sus componentes se pueda ajustar para adaptar el modelo probabilístico al problema a tratar y a la colección concreta, alcanzando la máxima efectividad recuperadora.

- Los resultados empíricos muestran que el rendimiento del S.R.I. es bastante dependiente de la colección utilizada. Fijado un modelo de red bayesiana documental y un conjunto de parámetros, la calidad de la recuperación varía con respecto al banco de pruebas con el que se experimente. Esto implica que no se pueda dar una combinación de valores óptima y común para todas las colecciones, siendo necesario adaptar el S.R.I. a la base de datos documental utilizada.

- La capacidad recuperadora se puede situar en los mismos niveles en los que se mueven otros modelos de recuperación basados en redes bayesianas y en niveles superiores con respecto a SMART.

- La incorporación de las relaciones existentes entre los términos de la colección es una herramienta que ofrece un incremento en la eficacia del S.R.I. El descubrimiento de dichas relaciones, bien directamente entre los términos, o vía documentos, origina un aumento de la expresividad, y por tanto, de la capacidad de ofrecer una información adicional muy útil a la hora de recuperar nuevos documentos relevantes. Éstos, de otra forma, permanecerían ocultos y no podrían ofrecerse al usuario. Los modelos de red aumentada, mixta y de subred extendida de documentos se configuran como los primeros modelos íntegros de recuperación de información basados en redes bayesianas que utilizan estas relaciones explícitamente.

Más específicamente, y agrupando las conclusiones por capítulos:

- *Sobre la expansión de consultas:*

  1. Utilizado el modelo para expandir consultas, se asegura un incremento de la capacidad recuperadora del S.R.I. con el que se esté trabajando.
  2. El uso del poliárbol de términos proporciona un conjunto de términos, los más relacionados con la consulta, que hace que ésta se vea completada de tal forma que se puedan recuperar documentos que son relevantes, pero que con la consulta original no se podrían obtener.
  3. La expansión se debe hacer con mucha cautela ya que la incorrecta utilización de los parámetros puede originar que la consulta expandida cambie el sentido original de la inicial.

- *Sobre el modelo de recuperación de red bayesiana documental:*

  1. La propagación exacta es inviable en este tipo de problemas, quedando totalmente descartada. Por otro lado, el uso de algoritmos aproximados para propagar en las redes bayesianas documentales no es aconsejable, al menos en lo que se refiere a los algoritmos de Monte Carlo actualmente existentes. La propagación exacta + evaluación es la adecuada para las topologías que manejamos, configurándose como un método eficiente, exacto y con buenos resultados.
  2. Las tres topologías para la subred de términos se presentan como un amplio abanico donde poder elegir aquélla que más se adecue a la colección con la que se esté trabajando.
  3. Los estimadores de las distribuciones de probabilidad marginal y condicionada sí ofrecen un comportamiento más homogéneo a lo largo de las distintas colecciones, mostrando así una independencia de las mismas. Por un lado, el estimador de las distribuciones de probabilidad marginal *pp2* es el más idóneo, indicando que es preferible utilizar una probabilidad a priori igual para todos los nodos sin padres. Por otro lado, el estimador de distribuciones de probabilidad condicionada con la que el S.R.I. alcanza una mayor capacidad recuperadora es *pc-J*.

4. En la red simple, la función de probabilidad que destaca, en cuanto a rendimiento, con respecto a las demás diseñadas, es la *fp10*.
5. La diferencia de las probabilidades a posteriori y a priori del documento para generar la ordenación de documentos es una técnica cuyo puesta en práctica parece adecuada para todas las colecciones. La excepción es CACM: tiene un número medio de términos por documento mucho menor que el resto de las colecciones y un número total de términos alto. Esto segundo implica que las probabilidades a priori y a posteriori de los términos sean muy similares y al realizar la diferencia, en muchos casos se anulan. Este hecho, unido a la primera característica citada, origina que sean pocos los términos que finalmente intervengan en la evaluación de las funciones de probabilidad en la subred de documentos, razón por la cual no es interesante utilizar esta técnica con este tipo de colecciones.
6. En cuanto a las dos técnicas diseñadas para incorporar la importancia de los términos de la consulta en la red bayesiana documental, el uso de la replicación de los nodos término que aparecen en un documento y a la vez en la consulta parece depender de las colecciones y es totalmente independiente de la topología usada. Esta dependencia de la colección también se mantiene cuando se emplean las evidencias parciales.
7. La selección de términos que se lleva a cabo para construir la red mixta, aunque no ofrezca los mejores resultados, es una técnica necesaria para colecciones con un gran número de documentos (y, consecuentemente, términos), reduciendo así las "dimensiones" del poliárbol.

- *Sobre la realimentación de la relevancia:*

1. La capacidad recuperadora del modelo de red bayesiana documental aumenta sensiblemente cuando se aplican los diferentes métodos de realimentación desarrollados, ofreciendo unos buenos resultados.
2. Las dos técnicas expuestas ofrecen resultados totalmente análogos.
3. El rendimiento de estas técnicas alcanza un nivel recuperador comparable con los conseguidos por otros modelos basados en redes bayesianas y otros clásicos.
4. La realimentación con el modelo de red bayesiana documental tiene una clara dependencia con respecto a la colección utilizada. Aunque siempre resulta beneficioso, la cuantía de la mejora varía considerablemente entre colecciones, al igual que ocurre con otros modelos de realimentación.
5. La metodología desarrollada para el modelo de red bayesiana documental incluye el hecho novedoso, con respecto a los modelos también basados en redes bayesianas, de la expansión negativa. La aportación de este tipo de expansión se ha mostrado de una gran utilidad para el rendimiento final de la técnica.
6. La incorporación de la calidad de la consulta (combinación convexa) a la realimentación basada en documentos es bastante útil. El comportamiento "inteligente"

de esta técnica hace que la expansión se potencie en mayor o menor grado según el rendimiento alcanzado en la primera consulta.

- *Sobre la extensión de la subred de documentos:*

    1. La extracción de las relaciones entre documentos en la subred de documentos se puede interpretar como una manera de establecer relaciones entre términos de forma indirecta. El motivo es que la única alternativa existente para relacionar documentos es a través de los términos que contienen [4].
    2. El rendimiento del S.R.I. en el que se han incluido las relaciones entre documentos tiene un comportamiento irregular, dependiente, en principio, de la colección y de la topología de la subred de términos. En cualquier caso, se abre un campo de estudio prometedor.

       Se ha observado que en aquellas colecciones donde se ha experimentado con los modelos aumentado y mixto (las relaciones entre términos se explicitan en el poliárbol aprendido), el rendimiento decrece considerablemente. Sin embargo, en el modelo simple, el comportamiento es justo al contrario: la eficacia recuperadora aumenta de forma apreciable. Esta conducta puede ser debida al hecho de que en los dos primeros modelos, al ampliar la subred de documentos, se produce una redundancia con respecto a las relaciones entre términos, ya que éstas figuran dos veces en la red bayesiana documental: una en la subred de términos y otra en la de documentos. Por tanto, la red simple, donde no se da esta circunstancia, más la nueva capa de documentos se configura como una alternativa eficiente (la propagación se reduce a la evaluación de funciones de probabilidad en las dos capas de documentos) e interesante por su alta calidad recuperadora, objeto de estudios futuros.
    3. La incorporación de una segunda capa de documentos mantiene intacto el modelo de red bayesiana documental subyacente a la subred de documentos (lo cual permite que éste pueda usarse con o sin extensión de la subred de documentos). Además y, como característica más importante, hace que la inferencia se pueda llevar a cabo por medio de la evaluación de funciones de probabilidad, lo cual garantiza, en ciertos casos, que la propagación sea exacta.

6.2. Trabajos futuros.

Son varias las líneas abiertas donde poder continuar la investigación en este campo. Seguidamente las vamos a exponer desde dos perspectivas diferentes:

[4] Otras formas posibles de relacionar directamente documentos entre sí, con independencia de los términos que contienen, no se han considerado en esta memoria. Estos métodos requerirían disponer de más información: por ejemplo, el uso de citas o referencias cruzadas entre documentos, o la evidencia acumulada en el tiempo de que ciertos documentos parecen ser relevantes conjuntamente para ciertas consultas.

1. *Desde el punto de vista técnico:*

    - Estudio detallado de las colecciones e identificación de patrones comunes.

      Una de las primeras tareas que se deberá hacer es estudiar detalladamente las cinco colecciones estándar de prueba y los resultados de la experimentación realizada, con objeto de identificar sus características e intentar establecer patrones claros. El objetivo, dada esa tipología, será determinar el conjunto de técnicas y parámetros con los que el modelo de recuperación de la red bayesiana documental alcanza un mayor rendimiento. Esto nos permitiría, si se deseara aplicar el S.R.I. a alguna nueva colección, tras realizar un análisis de la misma, conocer inmediatamente el conjunto idóneo de parámetros, según el tipo de base de datos documental, para el cual el S.R.I. rindiera de forma óptima.

      Un claro ejemplo de esta idea es el hecho de utilizar la diferencia de probabilidades a posteriori y a priori de un documento para establecer su posición en la ordenación final. Se ha observado, como hemos dicho antes, que con CACM esta técnica no es útil, al contrario que con el resto. Observando los estadísticos de esa colección se nota cómo el número medio de términos por documento es bajo. La razón es que CACM es una colección de resúmenes de artículos (abstracts) y no de textos completos. Este hecho nos serviría para evitar el uso de la diferencia de probabilidades en este tipo de colecciones.

    - Áreas de trabajo en el campo de la mejora de la estructura de la red y de algoritmos de propagación:

        - Subred de términos.
          Con respecto a la subred de términos, la idea básica será la de modificarla de modo que se permitan estructuras más potentes y, a la vez, versátiles. Una primera aproximación es poner en práctica un proceso de agrupamiento de términos, de tal forma que se creen un conjunto de clases donde quedarán agrupados. En cada clase, teniendo en cuenta que el tamaño de la misma será relativamente pequeño en comparación con el tamaño total del glosario, se puede aprender un G.D.A. más complejo que un poliárbol. Así, existirán varios grafos de dimensiones pequeñas. La propagación se llevará a cabo exclusivamente en aquellos G.D.A. en los que haya términos de la consulta, y aunque ésta tarea sea más costosa debido al tipo de grafo más complejo, al ser cada grafo pequeño, el costo será muy bajo. Además, existiría la posibilidad de lanzar paralelamente la propagación en cada grafo.
          Alternativamente, se puede pensar en el aprendizaje de un único grafo, pero más complejo que un poliárbol. En este caso, se deberán desarrollar nuevos métodos de aprendizaje y, posteriormente, de propagación para abordar estas tareas en un tiempo razonable. Uno de estos métodos será realizar una selección de características previa que permita determinar el conjunto de términos que realmente sean útiles para la recuperación y trabajar únicamente con ellos, consiguiendo así una reducción en la dimensión del problema.

En cuanto a la propagación, una alternativa puede ser realizar una propagación local, es decir, limitarla a un número de nodos concreto, a modo de radio prefijado, a partir de los nodos evidencia. Con esto evitamos la inferencia en la red completa, reduciendo así el tiempo de recuperación.

- Subred de documentos.
  Centrándonos en la subred de documentos, la tarea prioritaria será explotar la combinación red simple y subred extendida de documentos, mejorando la función de probabilidad existente y desarrollando nuevas para la nueva capa de documentos y nuevas formas conseguir las relaciones entre documentos (sobre todo para usarlas con la red aumentada y mixta). Esta información podría venir a partir de citas entre documentos, por ejemplo.
  Otro proceso alternativo será la creación de clases de documentos y el posterior aprendizaje de un grafo más complejo en cada una de ellas, de manera análoga a lo comentado en la subred de términos. De esta forma, las relaciones entre documentos se expresarían de una manera más rica. Además, se podrían aplicar algoritmos de propagación exacta, sin aumentar de manera considerable el tiempo de respuesta del S.R.I.

- Mejora de los métodos de realimentación de relevancia.

  Bajo este epígrafe incluiríamos el trabajo que pretende desarrollar métodos de realimentación de relevancia más potentes. Así, por un lado, buscaremos nuevas formas de calcular los vectores λ para la técnica presentada en esta memoria basada en la utilización de evidencias parciales, y por otro, alternativas a esa técnica que permitan añadir a la red los juicios de relevancia de los usuarios.

  Hasta ahora, y desde el punto de vista de la realimentación de relevancia, no se puede premiar a los términos más prometedores y, por el contrario, sí penalizar a los más malos. Este agravio, originado por el uso de evidencias parciales, se puede resolver utilizando la técnica de replicación de nodos. De esta manera, al introducir copias de un nodo en la red se puede hacer que tome más o menos importancia a la hora de evaluar las funciones de probabilidad, recompensando la calidad del término.

- Aprendizaje incremental y eliminación de material inservible de la red documental.

  En un entorno real, los documentos de los que se dispone van incrementándose paulatinamente con el paso del tiempo: por ejemplo, los fondos bibliográficos de una biblioteca están continuamente creciendo por la adquisición de nuevos ejemplares, o todavía más evidente: la proliferación de nuevas páginas web en Internet. Son dos ejemplos claros en los que la base de datos documental se debe actualizar una vez que ha sido construida. En nuestro caso, la llegada de un nuevo documento supondría tener que aprender de nuevo, y completamente, las subredes de términos y documentos, lo que implica un gasto en tiempo muy grande. Es por esta razón que está claramente justificado el hecho de desarrollar técnicas que permitan insertar nuevos nodos en las redes ya creadas, con sus correspondientes relaciones con el resto de nodos de las redes, realizando sólo modificaciones locales en dichas redes y sin que haya que aprenderlas completamente de nuevo.

Otro hecho muy común, más en Internet que en una biblioteca, es la desaparición de documentos. En este primer ámbito, el espacio de páginas web es totalmente dinámico, apareciendo y desapareciendo material HTML casi constantemente. Al igual que se debe reflejar la inserción de nuevo material en el G.D.A., se debe actualizar, eliminando el que ya no exista, para que sea totalmente consistente con respecto a la información que el usuario puede manejar.

- Otros usos de las redes bayesianas a la R.I.

  Algunos de estos usos podrían ser el desarrollo de nuevos métodos de clasificación documental de filtrado, empleando topologías de redes bayesianas específicas, así como el tema relacionado de la selección de características. También dentro del ámbito de la clasificación, los problemas específicos de filtrado de documentos.

2. *Desde el punto de vista del desarrollo de aplicaciones:*

   - Introducción de perfiles de usuario.

     Generalmente las consultas que hace un usuario a un S.R.I. versan más o menos sobre el mismo tema, por lo que parecería lógico ir creando, conforme se va utilizando el sistema, un perfil para cada usuario que vaya almacenando sus intereses a lo largo del tiempo. Este perfil serviría para limitar la búsqueda a un conjunto de documentos con los que está relacionado su perfil. De esta forma, se ahorraría mucho tiempo en las búsquedas y sólo se suministraría al usuario material sobre el que está realmente interesado.

     Una alternativa en la elaboración de los perfiles de usuario sería el uso de los juicios de relevancia que el usuario hace cuando aplica la técnica de realimentación.

   - Obviamente, para el desarrollo del trabajo experimental realizado en esta memoria, ha sido necesaria la creación de muy diversas herramientas software y la realización o modificación de otras. El resultado es un software bastante completo, pero muy heterogéneo, carente de un entorno común y de una interfaz mínimamente amigable. Pretendemos unificar todo este material software, para conseguir un paquete de recuperación de información completo y de libre disposición.

   - Experimentación con la colección TREC.

     Una de las tareas más atrayentes es, sin duda alguna, la evaluación del rendimiento de nuestro modelo sobre el banco de pruebas más difundido actualmente: las colecciones de documentos que pone a disposición de toda la comunidad investigadora la conferencia TREC. El atractivo de esta colección es doble para nosotros: por un lado, sus dimensiones son las de una colección real, poniendo a prueba nuestro S.R.I. para tratar este tipo de bases de datos documentales y, por otro, nos permitirá comprobar el rendimiento del modelo desarrollado con los conseguidos por una gran cantidad de sistemas, incluidos los basados en redes bayesianas.

   - Aplicación del S.R.I. a Internet como motor de búsqueda de páginas web.

Si la colección TREC es una de las más utilizadas cuando se está trabajando en S.R.I. experimentales, Internet y su casi infinito espacio de documentos web se constituye como el medio idóneo donde poner en manos del usuario nuestro S.R.I. basado en redes bayesianas de forma totalmente operativa. Así, el usuario podrá realizar búsquedas en la Red comparando la calidad de nuestro S.R.I. con los motores de búsqueda existentes actualmente.

A. Resultados del estudio estadístico de las colecciones estándar de prueba.

ADI	*Número Términos*	*idf medio*	*idf máximo*	*idf mínimo*
Media	25.4512	3.0207	4.4188	0.8725
Error típ. de la media	0.9131	2.366E-02	0.0	3.694E-02
Desviación típica	8.2689	0.2142	0.0	0.3345
Mínimo	12	2.32	4.42	0.63
Máximo	66	3.48	4.42	2.02
Percentil 25	20.7500	2.8766	4.4188	0.6347
Percentil 50 (Mediana)	24.0	3.0242	4.4188	0.6347
Percentil 75	29.0	3.1764	4.4188	0.9223

Cuadro A.1.: Estadísticos de los documentos de la colección ADI.

CACM	*Número Términos*	*idf medio*	*idf máximo*	*idf mínimo*
Media	23.2182	4.1989	7.5860	1.3567
Error típ. de la media	0.3517	9.862E-03	1.311E-02	1.166E-02
Desviación típica	19.9069	0.5582	0.7422	0.6599
Mínimo	2	2.29	3.39	0.88
Máximo	139	7.02	8.07	6.46
Percentil 25	6.0	3.8169	7.3793	0.8773
Percentil 50 (Mediana)	15.0	4.1107	8.0725	1.2362
Percentil 75	37.0	4.4699	8.0725	1.4794

Cuadro A.2.: Estadísticos de los documentos de la colección CACM.

CISI	*Número Términos*	*idf medio*	*idf máximo*	*idf mínimo*
Media	44.4199	3.2323	6.7221	0.4436
Error típ. de la media	0.4979	1.098E-02	2.238E-02	1.507E-03
Desviación típica	19.0252	0.4196	0.8550	5.759E-02
Mínimo	7	1.84	2.89	0.25
Máximo	157	5.57	7.29	0.76
Percentil 25	31.0	2.9740	6.1883	0.4081
Percentil 50 (Mediana)	42.0	3.1969	7.2869	0.4387
Percentil 75	55.0	3.4670	7.2869	0.4758

Cuadro A.3.: Estadísticos de los documentos de la colección CISI.

CRANFIELD	*Número Términos*	*idf medio*	*idf máximo*	*idf mínimo*
Media	52.9950	2.7935	6.5605	0.7469
Error típ. de la media	0.6009	9.823E-03	2.322E-02	4.348E-03
Desviación típica	22.4693	0.3673	0.8681	0.1626
Mínimo	10	1.75	2.81	0.65
Máximo	158	4.55	7.24	2.03
Percentil 25	37.0	2.5394	6.0744	0.6464
Percentil 50 (Mediana)	50.0	2.7496	7.2449	0.6464
Percentil 75	66.0	3.0095	7.2449	0.7068

Cuadro A.4.: Estadísticos de los documentos de la colección CRANFIELD.

MEDLARS	*Número Términos*	*idf medio*	*idf máximo*	*idf mínimo*
Media	51.1452	3.7117	6.7976	1.2675
Error típ. de la media	0.7019	1.035E-02	1.185E-02	6.945E-03
Desviación típica	22.5582	0.3326	0.3809	0.2232
Mínimo	8	2.79	4.46	1.06
Máximo	186	5.20	6.94	2.34
Percentil 25	35.0	3.4827	6.9412	1.0635
Percentil 50 (Mediana)	47.0	3.6767	6.9412	1.2308
Percentil 75	65.0	3.9077	6.9412	1.4039

Cuadro A.5.: Estadísticos de los documentos de la colección MEDLARS.

ADI	*Número Términos*	*idf medio*	*idf máximo*	*idf mínimo*
Media	6.8571	2.2630	3.8023	0.8119
Error típ. de la media	0.4604	8.340E-02	0.1192	5.627E-02
Desviación típica	2.7240	0.4934	0.7051	0.3329
Mínimo	3	1.04	1.47	0.63
Máximo	15	3.45	4.42	1.85
Percentil 25	5.0	1.9695	3.7257	0.6347
Percentil 50 (Mediana)	7.0	2.2153	3.7257	0.6347
Percentil 75	9.0	2.6667	4.4188	0.9223

Cuadro A.6.: Estadísticos de las consultas de la colección ADI.

CACM	*Número Términos*	*idf medio*	*idf máximo*	*idf mínimo*
Media	10.5938	3.7811	6.4378	1.6205
Error típ. de la media	0.7889	0.1035	0.2015	8.649E-02
Desviación típica	6.3113	0.8281	1.6123	0.6920
Mínimo	2	1.93	3.00	0.88
Máximo	24	6.08	8.07	3.64
Percentil 25	5.0	3.2404	5.1558	0.8773
Percentil 50 (Mediana)	9.0	3.7598	6.9739	1.4794
Percentil 75	15.7500	4.2702	8.0725	2.0171

Cuadro A.7.: Estadísticos de las consultas de la colección CACM.

CISI	*Número Términos*	*idf medio*	*idf máximo*	*idf mínimo*
Media	27.6875	2.8191	5.6	0.5963
Error típ. de la media	1.8040	4.616E-02	0.1329	9.764E-03
Desviación típica	19.0915	0.4885	1.4063	0.1033
Mínimo	3	1.36	1.75	0.29
Máximo	87	3.67	7.29	0.78
Percentil 25	10.2500	2.5858	4.6663	0.5470
Percentil 50 (Mediana)	26.0	2.8866	6.1883	0.6106
Percentil 75	41.0	3.1554	7.2869	0.6675

Cuadro A.8.: Estadísticos de las consultas de la colección CISI.

CRANFIELD	*Número Términos*	*idf medio*	*idf máximo*	*idf máximo*
Media	9.0670	2.5208	4.4241	1.1133
Error típ. de la media	0.2099	3.670E-02	8.075E-02	3.104E-02
Desviación típica	3.1409	0.5493	1.2086	0.4645
Mínimo	3	1.40	1.82	0.65
Máximo	20	4.47	7.24	3.10
Percentil 25	7.0	2.1438	3.4383	0.7068
Percentil 50 (Mediana)	9.0	2.4654	4.4723	0.9552
Percentil 75	11.0	2.8194	5.0477	1.3602

Cuadro A.9.: Estadísticos de las consultas de la colección CRANFIELD.

MEDLARS	*Número Términos*	*idf medio*	*idf máximo*	*idf mínimo*
Media	9.9667	3.4350	5.1485	1.8622
Error típ. de la media	1.0801	8.088E-02	0.1833	0.1130
Desviación típica	5.9160	0.4430	1.0037	0.6187
Mínimo	2	2.53	3.33	1.06
Máximo	27	4.46	6.94	3.76
Percentil 25	6.0	3.1703	4.5015	1.4359
Percentil 50 (Mediana)	8.0	3.4519	4.9285	1.7539
Percentil 75	14.0	3.6765	5.9439	2.0617

Cuadro A.10.: Estadísticos de las consultas de la colección MEDLARS .

B. Resultados empíricos detallados con la red bayesiana documental.

Exp.	SMART	fp1, pp1	fp1, pp2	fp1, pp3	fp3, pp1	fp3, pp2	fp3, pp3
REL. REC.	91	34	87	87	69	82	79
EXHAUST.	PRECISIÓN						
0.00	0.6966	0.2358	0.5767	0.5580	0.4175	0.4982	0.4336
0.10	0.6847	0.2207	0.5767	0.5509	0.4175	0.4962	0.4096
0.20	0.6397	0.1680	0.5332	0.5140	0.4112	0.4460	0.3843
0.30	0.5908	0.1544	0.4829	0.4721	0.3535	0.4059	0.3327
0.40	0.5669	0.1247	0.4279	0.3995	0.2838	0.3340	0.2879
0.50	0.5324	0.1098	0.4140	0.3881	0.2750	0.3280	0.2837
0.60	0.3991	0.0986	0.3136	0.2778	0.2311	0.2534	0.2359
0.70	0.2805	0.0874	0.2534	0.1952	0.1743	0.2128	0.1945
0.80	0.2636	0.0864	0.2415	0.1829	0.1623	0.1956	0.1741
0.90	0.2400	0.0785	0.2191	0.1608	0.1382	0.1707	0.1562
1.00	0.2385	0.0783	0.2179	0.1592	0.1379	0.1699	0.1543
M. 11PTS	0.4706	0.1311	0.3870	0.3508	0.2729	0.3192	0.2770
%C. 11PTS		-72.14	-17.8	-23.9	-42.0	-32.2	-41.13
M. 3PTS	0.4786	0.1214	0.3962	0.3617	0.2828	0.3232	0.2807
%C. 3PTS		-74.6	-17.2	-24.4	-40.9	-32.5	-41.3
E. EX.	0.5964	0.1863	0.5787	0.5584	0.4430	0.5344	0.4891
P. EX.	0.1733	0.0648	0.1657	0.1657	0.1314	0.1562	0.1505

Cuadro B.1.: Batería I con la red simple de ADI (Propagación aproximada).

Exp.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
REL. REC.	91	26	55	82	53	50	47	52	54
EXHAUST.	PRECISIÓN								
0.00	0.7007	0.2185	0.3250	0.4983	0.3574	0.3048	0.3623	0.3083	0.2668
0.10	0.6911	0.2011	0.3127	0.4969	0.3447	0.3048	0.3562	0.3083	0.2668
0.20	0.6548	0.1458	0.2899	0.4365	0.3349	0.2844	0.3339	0.2813	0.2427
0.30	0.5806	0.1334	0.2667	0.3868	0.2869	0.2562	0.3145	0.2556	0.2321
0.40	0.5514	0.1034	0.2366	0.3269	0.2427	0.2125	0.2617	0.2092	0.2032
0.50	0.5177	0.0870	0.2322	0.3193	0.2396	0.2034	0.2537	0.2087	0.2004
0.60	0.4063	0.0791	0.1669	0.2450	0.1876	0.1595	0.1743	0.1398	0.1502
0.70	0.2979	0.0738	0.1329	0.1956	0.1404	0.1181	0.1263	0.1235	0.1243
0.80	0.2822	0.0732	0.1316	0.1845	0.1384	0.1150	0.1195	0.1211	0.1163
0.90	0.2486	0.0714	0.1213	0.1603	0.1297	0.1022	0.1137	0.1101	0.1041
1.00	0.2454	0.0712	0.1205	0.1584	0.1296	0.0997	0.1126	0.1091	0.1021
M. 11PTS	0.4706	0.1144	0.2124	0.3098	0.2302	0.1964	0.2299	0.1977	0.182
%C. 11PTS		-75.7	-54.9	-34.2	-51.1	-58.3	-51.2	-58.0	-61.2
M. 3PTS	0.4849	0.1020	0.2179	0.3134	0.2376	0.2010	0.2357	0.2037	0.1865
%C. 3PTS		-79.0	-55.1	-35.4	-51.0	-58.6	-51.4	-58.0	-61.5
E. EX.	0.6120	0.1321	0.3679	0.5189	0.3811	0.3384	0.3488	0.3480	0.3733
P. EX.	0.2719	0.0495	0.1137	0.1894	0.1263	0.0975	0.1043	0.1025	0.1068

Cuadro B.2.: Batería I para la red simple de ADI: propagación exacta + evaluación, *pp1*, sin qf.

EXP.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
REL. REC.	91	25	84	82	84	86	82	90	89
EXHAUST.	PRECISIÓN								
0.00	0.7007	0.2170	0.5561	0.4983	0.6359	0.5607	0.5467	0.6168	0.6646
0.10	0.6911	0.1995	0.5481	0.4969	0.6274	0.5607	0.5372	0.6139	0.6646
0.20	0.6548	0.1443	0.4943	0.4365	0.5931	0.5521	0.5246	0.5885	0.6398
0.30	0.5806	0.1321	0.4730	0.3868	0.5124	0.5281	0.4964	0.5517	0.5788
0.40	0.5514	0.1018	0.4146	0.3269	0.4831	0.4946	0.4333	0.4873	0.5129
0.50	0.5177	0.0854	0.4002	0.3193	0.4572	0.4749	0.3871	0.4786	0.5080
0.60	0.4063	0.0778	0.3332	0.2450	0.3794	0.3719	0.3230	0.3640	0.4029
0.70	0.2979	0.0726	0.2566	0.1956	0.2718	0.2537	0.2266	0.2863	0.2840
0.80	0.2822	0.0721	0.2462	0.1845	0.2549	0.2310	0.2110	0.2795	0.2675
0.90	0.2486	0.0704	0.2121	0.1603	0.2253	0.2062	0.1791	0.2321	0.2235
1.00	0.2454	0.0702	0.2073	0.1584	0.2221	0.2036	0.1773	0.2284	0.2215
M. 11PTS	0.4706	0.1130	0.3765	0.3098	0.4239	0.4034	0.3675	0.4297	0.4516
%C. 11PTS		-76.0	-20.0	-34.2	-9.9	-14.3	-21.9	-8.7	-4.0
M. 3PTS	0.4849	0.1006	0.3802	0.3134	0.4351	0.4193	0.3742	0.4489	0.4717
%C. 3PTS		-79.3	-21.6	-35.4	-10.3	-13.5	-22.8	-7.4	-2.7
E. EX.	0.6120	0.1226	0.5666	0.5189	0.5563	0.5693	0.5404	0.5815	0.5679
P. EX.	0.2719	0.0476	0.2282	0.1894	0.2501	0.2377	0.2084	0.2636	0.2623

Cuadro B.3.: Batería I para la red simple de ADI: propagación exacta + evaluación, *pp2*, sin qf.

EXP.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
REL. REC.	91	25	84	82	80	85	84	84	92
EXHAUST.	PRECISIÓN								
0.00	0.7007	0.2170	0.5418	0.4983	0.6401	0.5541	0.5445	0.6195	0.6806
0.10	0.6911	0.1996	0.5204	0.4969	0.6354	0.5500	0.5341	0.6147	0.6759
0.20	0.6548	0.1440	0.4624	0.4365	0.5949	0.5451	0.5246	0.5833	0.6206
0.30	0.5806	0.1321	0.4479	0.3868	0.5088	0.5122	0.4981	0.5441	0.5563
0.40	0.5514	0.1021	0.4118	0.3269	0.4550	0.4850	0.4245	0.4742	0.5140
0.50	0.5177	0.0857	0.3977	0.3193	0.4351	0.4636	0.3823	0.4656	0.5089
0.60	0.4063	0.0779	0.3265	0.2450	0.3678	0.3603	0.3136	0.3545	0.3904
0.70	0.2979	0.0727	0.2547	0.1956	0.2674	0.2402	0.2221	0.2797	0.2801
0.80	0.2822	0.0722	0.2443	0.1845	0.2475	0.2232	0.2115	0.2710	0.2654
0.90	0.2486	0.0705	0.2102	0.1603	0.2187	0.1980	0.1801	0.2284	0.2179
1.00	0.2454	0.0703	0.2063	0.1584	0.2160	0.1967	0.1785	0.2243	0.2161
M. 11PTS	0.4706	0.1131	0.3658	0.3098	0.4170	0.3935	0.3649	0.4236	0.4478
%C. 11PTS		-76.0	-22.3	-34.2	-11.4	-16.4	-22.5	-10.0	-4.8
M. 3PTS	0.4849	0.1006	0.3681	0.3134	0.4258	0.4107	0.3728	0.4400	0.4650
%C. 3PTS		-79.2	-24.1	-35.4	-12.2	-15.3	-23.1	-9.3	-4.1
E. EX.	0.6120	0.1226	0.5567	0.5189	0.5295	0.5606	0.5484	0.5491	0.5878
P. EX.	0.2719	0.0476	0.2268	0.1894	0.2381	0.2278	0.2101	0.2481	0.2628

Cuadro B.4.: Batería I para la red simple de ADI: propagación exacta + evaluación, *pp3*, sin qf.

EXP.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
REL. REC.	246	1	10	26	26	1	18	10	5
EXHAUST.	PRECISIÓN								
0.00	0.7375	0.0183	0.1503	0.1370	0.3430	0.0218	0.1343	0.1169	0.0347
0.10	0.6806	0.0138	0.0468	0.0933	0.1500	0.0149	0.1072	0.0430	0.0308
0.20	0.5742	0.0130	0.0362	0.0781	0.1167	0.0122	0.0731	0.0426	0.0257
0.30	0.4774	0.0112	0.0155	0.0725	0.0377	0.0100	0.0557	0.0117	0.0186
0.40	0.4168	0.0105	0.0140	0.0714	0.0238	0.0097	0.0477	0.0088	0.0158
0.50	0.3515	0.0102	0.0134	0.0463	0.0112	0.0093	0.0386	0.0086	0.0144
0.60	0.2808	0.0099	0.0112	0.0383	0.0104	0.0087	0.0349	0.0081	0.0119
0.70	0.2192	0.0092	0.0102	0.0293	0.0090	0.0081	0.0201	0.0076	0.0110
0.80	0.1793	0.0087	0.0090	0.0225	0.0082	0.0078	0.0158	0.0072	0.0090
0.90	0.1255	0.0079	0.0075	0.0175	0.0068	0.0070	0.0089	0.0066	0.0071
1.00	0.1020	0.0065	0.0063	0.0126	0.0061	0.0063	0.0077	0.0059	0.0064
M. 11PTS	0.3768	0.0108	0.0291	0.0563	0.0657	0.0105	0.0495	0.0243	0.0168
%C. 11PTS		-97.1	-92.3	-85.1	-82.6	-97.2	-86.9	-93.6	-95.5
E. 3PTS	0.3683	0.0106	0.0195	0.0490	0.0454	0.0098	0.0425	0.0195	0.0164
%C. 3PTS		-97.1	-94.7	-86.7	-87.7	-97.3	-88.5	-94.7	-95.6
E. EX.	0.4172	0.0020	0.0166	0.0513	0.0517	0.0018	0.0656	0.0167	0.0238
P. EX.	0.3726	0.0022	0.0222	0.0578	0.0578	0.0022	0.0400	0.0222	0.0111

Cuadro B.5.: Batería I para la red simple de CACM: propagación exacta + evaluación, *pp1*, sin qf.

EXP.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
REL. REC.	246	1	78	26	99	77	83	99	147
EXHAUST.	PRECISIÓN								
0.00	0.7375	0.0182	0.3668	0.1370	0.5765	0.4603	0.5046	0.4488	0.7735
0.10	0.6806	0.0137	0.2992	0.0933	0.4434	0.3546	0.3782	0.3855	0.7361
0.20	0.5742	0.0128	0.2601	0.0781	0.3504	0.2654	0.3120	0.3246	0.5857
0.30	0.4774	0.0110	0.2233	0.0725	0.2707	0.2164	0.2640	0.2986	0.4932
0.40	0.4168	0.0104	0.2023	0.0714	0.2161	0.1675	0.2165	0.2554	0.3933
0.50	0.3515	0.0101	0.1680	0.0463	0.1855	0.1519	0.1627	0.1877	0.3221
0.60	0.2808	0.0098	0.1242	0.0383	0.1343	0.1260	0.1161	0.1512	0.2364
0.70	0.2192	0.0091	0.0961	0.0293	0.1030	0.1013	0.0810	0.1187	0.1675
0.80	0.1793	0.0087	0.0688	0.0225	0.0796	0.0760	0.0623	0.0941	0.1262
0.90	0.1255	0.0078	0.0256	0.0175	0.0274	0.0317	0.0291	0.0388	0.0553
1.00	0.1020	0.0064	0.0190	0.0126	0.0177	0.0203	0.0231	0.0267	0.0406
E. 11PTS	0.3768	0.0107	0.1685	0.0563	0.2186	0.1792	0.1954	0.2118	0.3573
%C. 11PTS		-97.2	-55.3	-85.1	-42.0	-52.4	-48.1	-43.8	-5.2
M. 3PTS	0.3683	0.0105	0.1656	0.0490	0.2052	0.1644	0.1790	0.2022	0.3447
%C. 3PTS		-97.1	-55.0	-86.7	-44.3	-55.4	-51.4	-45.1	-6.4
E. EX.	0.4172	0.0020	0.1842	0.0513	0.2545	0.1882	0.2449	0.2726	0.4177
P. EX.	0.3726	0.0022	0.1733	0.0578	0.2200	0.1711	0.1869	0.2210	0.3310

Cuadro B.6.: Batería I para la red simple de CACM: propagación exacta + evaluación, *pp2*, sin qf.

EXP.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
REL. REC.	246	1	79	26	98	77	82	100	146
EXHAUST.	PRECISIÓN								
0.00	0.7375	0.0182	0.3501	0.1370	0.5764	0.4610	0.5033	0.4495	0.7737
0.10	0.6806	0.0137	0.2946	0.0933	0.4441	0.3524	0.3767	0.3861	0.7396
0.20	0.5742	0.0128	0.2555	0.0781	0.3500	0.2639	0.3085	0.3225	0.5852
0.30	0.4774	0.0110	0.2232	0.0725	0.2706	0.2131	0.2577	0.2993	0.4950
0.40	0.4168	0.0104	0.2026	0.0714	0.2144	0.1653	0.2148	0.2553	0.3922
0.50	0.3515	0.0101	0.1693	0.0463	0.1832	0.1517	0.1608	0.1865	0.3210
0.60	0.2808	0.0098	0.1233	0.0383	0.1334	0.1261	0.1141	0.1508	0.2322
0.70	0.2192	0.0091	0.0946	0.0293	0.1022	0.0999	0.0795	0.1176	0.1676
0.80	0.1793	0.0087	0.0689	0.0225	0.0794	0.0755	0.0618	0.0930	0.1257
0.90	0.1255	0.0078	0.0252	0.0175	0.0267	0.0316	0.0284	0.0383	0.0548
1.00	0.1020	0.0064	0.0188	0.0126	0.0176	0.0200	0.0227	0.0268	0.0404
M. 11PTS	0.3768	0.0107	0.1660	0.0563	0.2180	0.1782	0.1935	0.2114	0.3570
%C. 11PTS		-97.2	-55.9	-85.1	-42.1	-52.7	-48.6	-43.9	-5.2
M. 3PTS	0.3683	0.0105	0.1645	0.0490	0.2042	0.1637	0.1771	0.2006	0.3440
%C. 3PTS		-97.1	-55.3	-86.7	-44.6	-55.6	-51.9	-45.5	-6.6
E. EX.	0.4172	0.0020	0.1875	0.0513	0.2534	0.1882	0.2440	0.2759	0.4159
P. EX.	0.3726	0.0022	0.1756	0.0578	0.2178	0.1711	0.1846	0.2232	0.3288

Cuadro B.7.: Batería I para la red simple de CACM: propagación exacta + evaluación, *pp3*, sin qf.

EXP.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
REL. REC.	343	35	155	142	165	67	124	103	77
EXHAUST.	PRECISIÓN								
0.00	0.6129	0.1255	0.3523	0.4052	0.4194	0.1280	0.2778	0.2951	0.1511
0.10	0.4485	0.0493	0.1899	0.1648	0.2159	0.0903	0.1741	0.1533	0.1074
0.20	0.3696	0.0433	0.1302	0.1283	0.1454	0.0778	0.1229	0.1014	0.0861
0.30	0.2950	0.0413	0.0971	0.1125	0.1072	0.0666	0.0967	0.0770	0.0726
0.40	0.2491	0.0400	0.0851	0.0935	0.0888	0.0595	0.0793	0.0614	0.0646
0.50	0.2120	0.0394	0.0740	0.0824	0.0768	0.0545	0.0704	0.0572	0.0578
0.60	0.1766	0.0384	0.0609	0.0718	0.0588	0.0487	0.0622	0.0511	0.0510
0.70	0.1285	0.0378	0.0526	0.0628	0.0477	0.0438	0.0499	0.0438	0.0453
0.80	0.0999	0.0372	0.0468	0.0549	0.0413	0.0402	0.0424	0.0401	0.0423
0.90	0.0689	0.0361	0.0383	0.0453	0.0348	0.0354	0.0366	0.0358	0.0358
1.00	0.0444	0.0352	0.0315	0.0341	0.0305	0.0310	0.0321	0.0305	0.0314
E. 11PTS	0.2459	0.0476	0.1053	0.1141	0.1151	0.0614	0.0949	0.0861	0.0678
%C. 11PTS		-80.7	-57.2	-53.6	-53.2	-75.0	-61.4	-65.0	-72.4
M. 3PTS	0.2272	0.0400	0.0837	0.0885	0.0878	0.0575	0.0786	0.0663	0.0620
%C. 3PTS		-82.4	-63.2	-61.0	-61.3	-74.7	-65.4	-70.8	-72.7
E. EX.	0.1810	0.0221	0.0727	0.0578	0.0786	0.0222	0.0610	0.0556	0.0333
P. EX.	0.3066	0.0364	0.1360	0.1246	0.1447	0.0588	0.1088	0.0904	0.0675

Cuadro B.8.: Batería I para la red simple de CISI: propagación exacta + evaluación, *pp1*, sin qf.

EXP.	SMART	fp1	fp2	fp3	fp4	fp5	fp6	fp8	fp10
REL. REC.	343	35	211	142	231	210	223	217	297
EXHAUST.	PRECISIÓN								
0.00	0.6129	0.1172	0.4993	0.4052	0.4783	0.4200	0.4728	0.4964	0.6026
0.10	0.4485	0.0491	0.2562	0.1648	0.2994	0.2247	0.2948	0.2697	0.3861
0.20	0.3696	0.0431	0.1821	0.1283	0.2194	0.1812	0.2072	0.2052	0.2919
0.30	0.2950	0.0411	0.1457	0.1125	0.1700	0.1473	0.1633	0.1612	0.2224
0.40	0.2491	0.0398	0.1189	0.0935	0.1311	0.1164	0.1347	0.1212	0.1777
0.50	0.2120	0.0393	0.1058	0.0824	0.1072	0.0946	0.1087	0.1056	0.1479
0.60	0.1766	0.0383	0.0844	0.0718	0.0837	0.0763	0.0883	0.0869	0.1183
0.70	0.1285	0.0378	0.0665	0.0628	0.0645	0.0636	0.0706	0.0696	0.0892
0.80	0.0999	0.0371	0.0563	0.0549	0.0513	0.0522	0.0545	0.0590	0.0733
0.90	0.0689	0.0361	0.0427	0.0453	0.0389	0.0422	0.0442	0.0429	0.0489
1.00	0.0444	0.0352	0.0332	0.0341	0.0326	0.0330	0.0371	0.0334	0.0382
M. 11PTS	0.2459	0.0467	0.1447	0.1141	0.1524	0.1319	0.1524	0.1501	0.1997
%C. 11PTS		-81.0	-41.2	-53.6	-38.0	-46.3	-38.0	-39.0	-18.8
M. 3PTS	0.2272	0.0398	0.1147	0.0885	0.1260	0.1093	0.1235	0.1233	0.1710
%C. 3PTS		-82.5	-49.5	-61.0	-44.5	-51.9	-45.6	-45.7	-24.7
E. EX.	0.1810	0.0221	0.0952	0.0578	0.1011	0.0914	0.1105	0.1053	0.1537
P. EX.	0.3066	0.0364	0.1851	0.1246	0.2026	0.1842	0.1962	0.1904	0.2618

Cuadro B.9.: Batería I para la red simple de CISI: propagación exacta + evaluación, *pp2*, sin qf.

EXP.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
REL. REC.	343	35	210	142	230	208	221	217	294
EXHAUST.	PRECISIÓN								
0.00	0.6129	0.1239	0.4832	0.4052	0.4849	0.4198	0.4718	0.4997	0.6009
0.10	0.4485	0.0491	0.2536	0.1648	0.2990	0.2252	0.2945	0.2649	0.3889
0.20	0.3696	0.0431	0.1791	0.1283	0.2181	0.1801	0.2051	0.2039	0.2905
0.30	0.2950	0.0411	0.1434	0.1125	0.1684	0.1465	0.1616	0.1588	0.2206
0.40	0.2491	0.0398	0.1173	0.0935	0.1302	0.1138	0.1340	0.1198	0.1757
0.50	0.2120	0.0393	0.1036	0.0824	0.1071	0.0935	0.1084	0.1039	0.1480
0.60	0.1766	0.0383	0.0846	0.0718	0.0836	0.0761	0.0878	0.0858	0.1167
0.70	0.1285	0.0378	0.0665	0.0628	0.0640	0.0634	0.0703	0.0692	0.0880
0.80	0.0999	0.0371	0.0562	0.0549	0.0512	0.0523	0.0545	0.0588	0.0731
0.90	0.0689	0.0361	0.0427	0.0453	0.0390	0.0421	0.0440	0.0428	0.0488
1.00	0.0444	0.0352	0.0333	0.0341	0.0325	0.0330	0.0370	0.0334	0.0381
M. 11PTS	0.2459	0.0473	0.1421	0.1141	0.1525	0.1314	0.1517	0.1492	0.1990
%C. 11PTS		-80.7	-42.2	-53.6	-38.0	-46.6	-38.3	-39.3	-19.1
M. 3PTS	0.2272	0.0398	0.1130	0.0885	0.1254	0.1086	0.1227	0.1222	0.1705
%C. 3PTS		-82.5	-50.3	-61.0	-44.8	-52.2	-46.0	-46.2	-24.9
E. EX.	0.1810	0.0221	0.0950	0.0578	0.1008	0.0910	0.1094	0.1043	0.1530
P. EX.	0.3066	0.0364	0.1842	0.1246	0.2018	0.1825	0.1944	0.1904	0.2592

Cuadro B.10.: Batería II para la red simple de CISI: propagación exacta + evaluación, *pp3*, sin qf.

EXP.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
REL. REC.	824	23	444	467	463	291	261	331	331
EXHAUST.	PRECISIÓN								
0.00	0.8020	0.0260	0.5807	0.5305	0.6222	0.3324	0.2898	0.4474	0.2233
0.10	0.7780	0.0197	0.5339	0.4967	0.5726	0.3061	0.2491	0.4093	0.2168
0.20	0.6651	0.0107	0.4270	0.4012	0.4491	0.2179	0.1766	0.3201	0.1863
0.30	0.5598	0.0091	0.3014	0.2972	0.3199	0.1495	0.1336	0.2065	0.1444
0.40	0.4656	0.0084	0.2264	0.2274	0.2345	0.1060	0.0963	0.1531	0.1162
0.50	0.4131	0.0079	0.1944	0.1946	0.1881	0.0874	0.0822	0.1240	0.0988
0.60	0.3195	0.0075	0.1215	0.1377	0.1306	0.0594	0.0606	0.0848	0.0776
0.70	0.2467	0.0072	0.0776	0.0939	0.0832	0.0382	0.0421	0.0465	0.0546
0.80	0.2024	0.0070	0.0578	0.0749	0.0687	0.0279	0.0291	0.0359	0.0433
0.90	0.1445	0.0067	0.0367	0.0536	0.0427	0.0209	0.0199	0.0258	0.0295
1.00	0.1267	0.0066	0.0326	0.0494	0.0400	0.0197	0.0177	0.0234	0.0271
M. 11PTS	0.4294	0.0106	0.2355	0.2325	0.2501	0.1241	0.1088	0.1706	0.1107
%C. 11PTS		-97.5	-45.2	-45.9	-41.7	-71.1	-74.7	-60.3	-74.2
M. 3PTS	0.4269	0.0085	0.2264	0.2236	0.2353	0.1111	0.0960	0.1600	0.1095
%C. 3PTS		-98.0	-47.0	-47.6	-44.9	-74.0	-77.5	-62.5	-74.4
E. EX.	0.5137	0.0093	0.2924	0.3079	0.2958	0.1900	0.1751	0.2262	0.2193
P. EX.	0.2850	0.0068	0.1358	0.1468	0.1474	0.0865	0.0773	0.0997	0.1006

Cuadro B.11.: Batería I para la red simple de CRANFIELD: propagación exacta + evaluación, *pp1*, sin qf.

EXP.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
REL. REC.	824	23	630	467	656	647	604	621	777
EXHAUST.	PRECISIÓN								
0.00	0.8020	0.0259	0.7042	0.5305	0.6952	0.7339	0.5921	0.6835	0.7961
0.10	0.7780	0.0195	0.6795	0.4967	0.6677	0.6872	0.5575	0.6447	0.7774
0.20	0.6651	0.0106	0.5419	0.4012	0.5478	0.5587	0.4564	0.5255	0.6547
0.30	0.5598	0.0090	0.4221	0.2972	0.4216	0.4440	0.3559	0.4109	0.5432
0.40	0.4656	0.0083	0.3510	0.2274	0.3266	0.3546	0.2822	0.3522	0.4390
0.50	0.4131	0.0078	0.2963	0.1946	0.2771	0.2975	0.2459	0.2973	0.3856
0.60	0.3195	0.0075	0.2064	0.1377	0.2140	0.2106	0.1876	0.2144	0.2940
0.70	0.2467	0.0071	0.1523	0.0939	0.1467	0.1492	0.1216	0.1525	0.2094
0.80	0.2024	0.0069	0.1217	0.0749	0.1172	0.1195	0.0852	0.1171	0.1689
0.90	0.1445	0.0067	0.0787	0.0536	0.0738	0.0861	0.0549	0.0744	0.1126
1.00	0.1267	0.0066	0.0704	0.0494	0.0664	0.0780	0.0481	0.0682	0.0965
M. 11PTS	0.4294	0.0105	0.3295	0.2325	0.3231	0.3381	0.2716	0.3219	0.4070
%C. 11PTS		-97.5	-23.3	-45.9	-24.8	-21.3	-36.8	-25.0	-5.2
M. 3PTS	0.4269	0.0085	0.3200	0.2236	0.3141	0.3252	0.2625	0.3133	0.4030
%C. 3PTS		-98.0	-25.0	-47.6	-26.4	-23.8	-38.5	-26.6	-5.6
E. EX.	0.5137	0.0093	0.4030	0.3079	0.4059	0.4129	0.3770	0.3966	0.4762
P. EX.	0.2850	0.0068	0.2078	0.1468	0.2133	0.2148	0.1904	0.2031	0.2614

Cuadro B.12.: Batería I para la red simple de CRANFIELD: propagación exacta + evaluación, *pp2*, sin qf.

EXP.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
REL. REC.	824	23	625	467	655	634	596	620	776
EXHAUST.	PRECISIÓN								
0.00	0.8020	0.0259	0.7013	0.5305	0.6900	0.7238	0.5869	0.6824	0.7933
0.10	0.7780	0.0195	0.6768	0.4967	0.6625	0.6763	0.5533	0.6432	0.7716
0.20	0.6651	0.0106	0.5404	0.4012	0.5432	0.5477	0.4512	0.5243	0.6553
0.30	0.5598	0.0090	0.4127	0.2972	0.4215	0.4356	0.3517	0.4085	0.5430
0.40	0.4656	0.0083	0.3425	0.2274	0.3273	0.3464	0.2781	0.3510	0.4340
0.50	0.4131	0.0078	0.2932	0.1946	0.2747	0.2912	0.2427	0.2950	0.3825
0.60	0.3195	0.0075	0.2023	0.1377	0.2103	0.2023	0.1838	0.2120	0.2884
0.70	0.2467	0.0071	0.1465	0.0939	0.1437	0.1427	0.1193	0.1508	0.2069
0.80	0.2024	0.0069	0.1179	0.0749	0.1155	0.1133	0.0830	0.1151	0.1664
0.90	0.1445	0.0067	0.0751	0.0536	0.0720	0.0826	0.0536	0.0729	0.1120
1.00	0.1267	0.0066	0.0672	0.0494	0.0645	0.0745	0.0471	0.0668	0.0966
M. 11PTS	0.4294	0.0105	0.3251	0.2325	0.3205	0.3306	0.2683	0.3202	0.4045
%C. 11PTS		-97.5	-24.3	-45.9	-25.4	-23.0	-37.5	-25.4	-5.8
M. 3PTS	0.4269	0.0085	0.3171	0.2236	0.3111	0.3174	0.2590	0.3115	0.4014
%C. 3PTS		-98.0	-25.7	-47.6	-27.1	-25.6	-39.3	-27.0	-6.0
E. EX.	0.5137	0.0093	0.4002	0.3079	0.4058	0.4054	0.3730	0.3964	0.4735
P. EX.	0.2850	0.0068	0.2043	0.1468	0.2118	0.2091	0.1880	0.2028	0.2619

Cuadro B.13.: Batería I para la red simple de CRANFIELD: propagación exacta + evaluación, *pp3*, sin qf.

EXP.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
REL. REC.	260	8	155	196	132	100	89	127	138
EXHAUST.	PRECISIÓN								
0.00	0.9137	0.0663	0.7070	0.8679	0.7452	0.4160	0.4942	0.6441	0.5726
0.10	0.8181	0.0400	0.5818	0.6876	0.5520	0.3456	0.3873	0.4962	0.5047
0.20	0.7512	0.0307	0.4265	0.5635	0.4058	0.2977	0.2580	0.3257	0.3963
0.30	0.6858	0.0287	0.3674	0.4726	0.3138	0.2386	0.2084	0.2795	0.2864
0.40	0.6344	0.0277	0.2546	0.4205	0.2048	0.1897	0.1302	0.1891	0.2446
0.50	0.5543	0.0268	0.2129	0.3580	0.1627	0.1420	0.1092	0.1527	0.1809
0.60	0.4789	0.0255	0.1662	0.3180	0.1204	0.1144	0.0861	0.1212	0.1399
0.70	0.4297	0.0249	0.1366	0.2455	0.0915	0.0937	0.0705	0.1016	0.0980
0.80	0.3527	0.0240	0.0930	0.1920	0.0612	0.0757	0.0483	0.0725	0.0730
0.90	0.2527	0.0235	0.0658	0.1350	0.0462	0.0589	0.0369	0.0523	0.0458
1.00	0.1193	0.0232	0.0393	0.0662	0.0308	0.0387	0.0314	0.0337	0.0334
M. 11PTS	0.5446	0.0310	0.2774	0.3934	0.2486	0.1828	0.1691	0.2244	0.2342
%C. 11PTS		-94.3	-49.1	-27.8	-54.4	-66.4	-68.9	-58.8	-57.0
M. 3PTS	0.5527	0.0272	0.2442	0.3712	0.2099	0.1718	0.1385	0.1836	0.2168
%C. 3PTS		-95.1	-55.8	-32.8	-62.0	-68.9	-74.9	-66.8	-60.8
E. EX.	0.4104	0.0119	0.2468	0.3095	0.2118	0.1586	0.1495	0.2033	0.2199
P. EX.	0.5778	0.0178	0.3444	0.4356	0.2933	0.2222	0.1978	0.2822	0.3067

Cuadro B.14.: Batería I para la red simple de MEDLARS: propagación exacta + evaluación, *pp1*, sin qf.

EXP.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
REL. REC.	260	6	242	196	225	236	196	242	266
EXHAUST.	PRECISIÓN								
0.00	0.9137	0.0807	0.9628	0.8679	0.9262	0.9110	0.7622	0.9060	0.9403
0.10	0.8181	0.0400	0.8620	0.6876	0.8172	0.8272	0.7146	0.8403	0.8559
0.20	0.7512	0.0306	0.7161	0.5635	0.6844	0.7215	0.6089	0.7022	0.7572
0.30	0.6858	0.0286	0.6376	0.4726	0.5759	0.6385	0.5200	0.6481	0.7057
0.40	0.6344	0.0277	0.5301	0.4205	0.4630	0.5613	0.4001	0.5383	0.6382
0.50	0.5543	0.0267	0.4397	0.3580	0.3899	0.4438	0.3389	0.4609	0.5453
0.60	0.4789	0.0255	0.3858	0.3180	0.3165	0.3682	0.2491	0.4116	0.4651
0.70	0.4297	0.0249	0.3173	0.2455	0.2474	0.3080	0.1881	0.3423	0.3761
0.80	0.3527	0.0240	0.2348	0.1920	0.1881	0.2385	0.1240	0.2614	0.2990
0.90	0.2527	0.0234	0.1613	0.1350	0.1199	0.1352	0.0924	0.1671	0.1996
1.00	0.1193	0.0232	0.0856	0.0662	0.0686	0.0766	0.0571	0.0872	0.0976
M. 11PTS	0.5446	0.0323	0.4848	0.3934	0.4361	0.4754	0.3687	0.4878	0.5345
%C. 11PTS		-94.1	-11.0	-27.8	-19.9	-12.7	-32.3	-10.4	-1.8
M. 3PTS	0.5527	0.0271	0.4635	0.3712	0.4208	0.4680	0.3572	0.4748	0.5338
%C. 3PTS		-95.1	-16.1	-32.8	-23.9	-15.3	-35.4	-14.1	-3.4
E. EX.	0.4104	0.0096	0.3767	0.3095	0.3487	0.3686	0.3096	0.3773	0.4152
P. EX.	0.5778	0.0133	0.5378	0.4356	0.5000	0.5244	0.4356	0.5378	0.5911

Cuadro B.15.: Batería I para la red simple de MEDLARS: propagación exacta + evaluación, *pp2*, sin qf.

EXP.	SMART	fp1-	fp2-	fp3-	fp4-	fp5-	fp6-	fp8-	fp10-
REL. REC.	260	6	241	196	225	237	197	241	266
EXHAUST.	PRECISIÓN								
0.00	0.9137	0.0807	0.9619	0.8679	0.9262	0.9111	0.7622	0.9049	0.9237
0.10	0.8181	0.0400	0.8611	0.6876	0.8172	0.8314	0.7146	0.8392	0.8559
0.20	0.7512	0.0306	0.7131	0.5635	0.6855	0.7209	0.6037	0.7030	0.7564
0.30	0.6858	0.0286	0.6342	0.4726	0.5745	0.6357	0.5188	0.6465	0.7053
0.40	0.6344	0.0277	0.5302	0.4205	0.4626	0.5620	0.3955	0.5352	0.6363
0.50	0.5543	0.0267	0.4382	0.3580	0.3887	0.4405	0.3361	0.4598	0.5424
0.60	0.4789	0.0255	0.3863	0.3180	0.3142	0.3658	0.2463	0.4089	0.4658
0.70	0.4297	0.0249	0.3200	0.2455	0.2468	0.3074	0.1866	0.3428	0.3747
0.80	0.3527	0.0240	0.2331	0.1920	0.1867	0.2372	0.1225	0.2613	0.2969
0.90	0.2527	0.0234	0.1615	0.1350	0.1185	0.1331	0.0904	0.1646	0.1968
1.00	0.1193	0.0232	0.0848	0.0662	0.0680	0.0763	0.0564	0.0862	0.0966
M. 11PTS	0.5446	0.0323	0.4840	0.3934	0.4354	0.4747	0.3666	0.4866	0.5319
%C. 11PTS		-94.1	-11.1	-27.8	-20.1	-12.8	-32.7	-10.7	-2.3
M. 3PTS	0.5527	0.0271	0.4615	0.3712	0.4203	0.4662	0.3541	0.4747	0.5319
%C. 3PTS		-95.1	-16.5	-32.8	-24.0	-15.7	-35.9	-14.1	-3.8
E. EX.	0.4104	0.0096	0.3755	0.3095	0.3487	0.3686	0.3104	0.3764	0.4152
P. EX.	0.5778	0.0133	0.5356	0.4356	0.5000	0.5267	0.4378	0.5356	0.5911

Cuadro B.16.: Batería I para la red simple de MEDLARS: propagación exacta + evaluación, *pp3*, sin qf.

EXP.	SMART	0, –	0	1, –	1
REL. REC.	91	90	88	91	90
EXHAUST.	PRECISIÓN				
0.00	0.7007	0.6168	0.6403	0.6246	0.6541
0.10	0.6911	0.6139	0.6345	0.6161	0.6469
0.20	0.6548	0.5885	0.6060	0.5889	0.6176
0.30	0.5806	0.5517	0.5739	0.5533	0.5753
0.40	0.5514	0.4873	0.5153	0.5008	0.5238
0.50	0.5177	0.4786	0.4981	0.4904	0.5039
0.60	0.4063	0.3640	0.3827	0.3780	0.3925
0.70	0.2979	0.2863	0.2989	0.2854	0.2995
0.80	0.2822	0.2795	0.2886	0.2793	0.2840
0.90	0.2486	0.2321	0.2617	0.2394	0.2561
1.00	0.2454	0.2284	0.2590	0.2355	0.2536
M. 11PTS	0.4706	0.4297	0.4508	0.4356	0.4552
%C. 11PTS		-8.7	-4.2	-7.4	-3.3
M. 3PTS	0.4849	0.4489	0.4642	0.4529	0.4685
%C. 3PTS		-7.4	-4.3	-6.6	-3.4
E. EX.	0.6120	0.5815	0.5984	0.5934	0.6088
P. EX.	0.2719	0.2636	0.2736	0.2706	0.2723

Cuadro B.17.: Batería II para la red simple de ADI: *pp2* y *fp8*, con y sin *qf*, con y sin diferencia de probabilidades.

EXP.	SMART	0, –	0	1, –	1
REL. REC.	91	89	90	90	91
EXHAUST.	PRECISIÓN				
0.00	0.7007	0.6646	0.7030	0.6646	0.7007
0.10	0.6911	0.6646	0.6941	0.6589	0.6911
0.20	0.6548	0.6398	0.6557	0.6407	0.6548
0.30	0.5806	0.5788	0.5772	0.5776	0.5806
0.40	0.5514	0.5129	0.5414	0.5294	0.5514
0.50	0.5177	0.5080	0.5118	0.5194	0.5177
0.60	0.4063	0.4029	0.3978	0.4179	0.4065
0.70	0.2979	0.2840	0.3028	0.2903	0.2982
0.80	0.2822	0.2675	0.2890	0.2741	0.2827
0.90	0.2486	0.2235	0.2538	0.2306	0.2497
1.00	0.2454	0.2215	0.2506	0.2290	0.2469
M. 11PTS	0.4706	0.4516	0.4707	0.4575	0.4709
%C. 11PTS		-4.0	0.0	-2.8	0.1
M. 3PTS	0.4849	0.4717	0.4855	0.4780	0.4851
%C. 3PTS		-2.7	0.1	-1.4	0.0
E. EX.	0.6120	0.5679	0.6025	0.5688	0.6120
P. EX.	0.2719	0.2623	0.2737	0.2723	0.2719

Cuadro B.18.: Batería II para la red simple de ADI: *pp2* y *fp10*, con y sin *qf*, con y sin diferencia de probabilidades.

EXP.	SMART	0, –	0	1, –	1
REL. REC.	246	99	100	100	102
EXHAUST.	PRECISIÓN				
0.00	0.7375	0.4488	0.4644	0.4532	0.4701
0.10	0.6806	0.3855	0.3938	0.3900	0.3987
0.20	0.5742	0.3246	0.3335	0.3401	0.3449
0.30	0.4774	0.2986	0.2927	0.3132	0.3060
0.40	0.4168	0.2554	0.2522	0.2710	0.2691
0.50	0.3515	0.1877	0.1988	0.1987	0.2048
0.60	0.2808	0.1512	0.1574	0.1465	0.1565
0.70	0.2192	0.1187	0.1324	0.1173	0.1291
0.80	0.1793	0.0941	0.1044	0.0930	0.1007
0.90	0.1255	0.0388	0.0472	0.0413	0.0501
1.00	0.1020	0.0267	0.0319	0.0321	0.0351
M. 11PTS	0.3768	0.2118	0.2190	0.2179	0.2241
%C. 11PTS		-43.8	-41.9	-42.2	-40.5
M. 3PTS	0.3683	0.2022	0.2122	0.2106	0.2168
%C. 3PTS		-45.1	-42.4	-42.8	-41.1
E. EX.	0.4172	0.2726	0.2721	0.2702	0.2811
P. EX.	0.3726	0.2210	0.2222	0.2232	0.2267

Cuadro B.19.: Batería II para la red simple de CACM: *pp2* y *fp8*, con y sin *qf*, con y sin diferencia de probabilidades.

EXP.	SMART	0, –	0	1, –	1
REL. REC.	246	147	142	146	143
EXHAUST.	PRECISIÓN				
0.00	0.7375	0.7735	0.7759	0.7168	0.7295
0.10	0.6806	0.7361	0.7253	0.6741	0.6684
0.20	0.5742	0.5857	0.5812	0.5615	0.5415
0.30	0.4774	0.4932	0.4742	0.4571	0.4426
0.40	0.4168	0.3933	0.3888	0.3823	0.3787
0.50	0.3515	0.3221	0.3172	0.3222	0.3119
0.60	0.2808	0.2364	0.2442	0.2432	0.2462
0.70	0.2192	0.1675	0.1881	0.1831	0.1969
0.80	0.1793	0.1262	0.1420	0.1255	0.1437
0.90	0.1255	0.0553	0.0621	0.0587	0.0705
1.00	0.1020	0.0406	0.0417	0.0432	0.0485
M. 11PTS	0.3768	0.3573	0.3582	0.3425	0.3435
%C. 11PTS		-5.2	-4.9	-9.1	-8.8
M. 3PTS	0.3683	0.3447	0.3468	0.3364	0.3323
%C. 3PTS		-6.4	-5.8	-8.7	-9.8
E. EX.	0.4172	0.4177	0.3873	0.4131	0.4031
P. EX.	0.3726	0.3310	0.3166	0.3299	0.3232

Cuadro B.20.: Batería II para la red simple de CACM: *pp2* y *fp10*, con y sin *qf*, con y sin diferencia de probabilidades.

EXP.	SMART	0, –	0	1, –	1
REL. REC.	343	217	257	264	300
EXHAUST.	PRECISIÓN				
0.00	0.6129	0.4964	0.5461	0.5566	0.5873
0.10	0.4485	0.2697	0.3239	0.3549	0.3973
0.20	0.3696	0.2052	0.2477	0.2594	0.3031
0.30	0.2950	0.1612	0.2024	0.2157	0.2494
0.40	0.2491	0.1212	0.1605	0.1740	0.2056
0.50	0.2120	0.1056	0.1452	0.1572	0.1847
0.60	0.1766	0.0869	0.1224	0.1267	0.1565
0.70	0.1285	0.0696	0.0917	0.0855	0.1098
0.80	0.0999	0.0590	0.0749	0.0667	0.0829
0.90	0.0689	0.0429	0.0529	0.0492	0.0597
1.00	0.0444	0.0334	0.0375	0.0355	0.0394
M. 11PTS	0.2459	0.1501	0.1823	0.1892	0.2160
%C. 11PTS		-39.0	-25.9	-23.1	-12.2
M. 3PTS	0.2272	0.1233	0.1560	0.1611	0.1902
%C. 3PTS		-45.7	-31.3	-29.1	-16.3

Cuadro B.21.: Batería II para la red simple de CISI: *pp2* y *fp8*, con y sin *qf*, con y sin diferencia de probabilidades.

EXP.	SMART	0, –	0	1, –	1
REL. REC.	343	297	312	349	369
EXHAUST.	PRECISIÓN				
0.00	0.6129	0.6026	0.6224	0.6691	0.6689
0.10	0.4485	0.3861	0.3965	0.4719	0.4831
0.20	0.3696	0.2919	0.3173	0.3793	0.3971
0.30	0.2950	0.2224	0.2508	0.3020	0.3151
0.40	0.2491	0.1777	0.2055	0.2527	0.2674
0.50	0.2120	0.1479	0.1740	0.2129	0.2311
0.60	0.1766	0.1183	0.1478	0.1668	0.1911
0.70	0.1285	0.0892	0.1133	0.1224	0.1364
0.80	0.0999	0.0733	0.0894	0.0909	0.1010
0.90	0.0689	0.0489	0.0632	0.0606	0.0692
1.00	0.0444	0.0382	0.0466	0.0429	0.0461
M. 11PTS	0.2459	0.1997	0.2206	0.2520	0.2642
%C. 11PTS		-18.8	-10.3	2.5	7.4
M. 3PTS	0.2272	0.1710	0.1936	0.2277	0.2430
%C. 3PTS		-24.7	-14.8	0.2	7.0
E. EX.	0.1810	0.1537	0.1599	0.1794	0.1881
P. EX.	0.3066	0.2618	0.2794	0.3118	0.3294

Cuadro B.22.: Batería II para la red simple de CISI: *pp2* y *fp10*, con y sin *qf*, con y sin diferencia de probabilidades.

EXP.	SMART	0, –	0	1, –	1
REL. REC.	824	621	708	624	709
EXHAUST.	PRECISIÓN				
0.00	0.8020	0.6835	0.7256	0.6844	0.7244
0.10	0.7780	0.6447	0.6968	0.6463	0.6944
0.20	0.6651	0.5255	0.5858	0.5247	0.5872
0.30	0.5598	0.4109	0.4707	0.4144	0.4686
0.40	0.4656	0.3522	0.4014	0.3504	0.3981
0.50	0.4131	0.2973	0.3460	0.2963	0.3446
0.60	0.3195	0.2144	0.2685	0.2110	0.2649
0.70	0.2467	0.1525	0.2027	0.1530	0.2016
0.80	0.2024	0.1171	0.1633	0.1184	0.1638
0.90	0.1445	0.0744	0.1139	0.0762	0.1145
1.00	0.1267	0.0682	0.1020	0.0701	0.1029
M. 11PTS	0.4294	0.3219	0.3706	0.3223	0.3695
%C. 11PTS		-25.0	-13.7	-24.9	-13.9
M. 3PTS	0.4269	0.3133	0.3650	0.3131	0.3652
%C. 3PTS		-26.6	-14.5	-26.6	-14.4
E. EX.	0.5137	0.3966	0.4488	0.4003	0.4507
P. EX.	0.2850	0.2031	0.2378	0.2044	0.2387

Cuadro B.23.: Batería II para la red simple de CRANFIELD: *pp2* y *fp8*, con y sin *qf*, con y sin diferencia de probabilidades.

EXP.	SMART	0, –	0	1, –	1
REL. REC.	824	777	816	779	825
EXHAUST.	PRECISIÓN				
0.00	0.8020	0.7961	0.8046	0.7943	0.8024
0.10	0.7780	0.7774	0.7830	0.7745	0.7789
0.20	0.6651	0.6547	0.6660	0.6515	0.6658
0.30	0.5598	0.5432	0.5629	0.5413	0.5642
0.40	0.4656	0.4390	0.4687	0.4434	0.4659
0.50	0.4131	0.3856	0.4205	0.3845	0.4171
0.60	0.3195	0.2940	0.3287	0.2916	0.3213
0.70	0.2467	0.2094	0.2502	0.2106	0.2493
0.80	0.2024	0.1689	0.2015	0.1703	0.2031
0.90	0.1445	0.1126	0.1428	0.1152	0.1450
1.00	0.1267	0.0965	0.1261	0.0982	0.1274
M. 11PTS	0.4294	0.4070	0.4323	0.4069	0.4309
%C. 11PTS		-5.2	0.7	-5.3	0.4
M. 3PTS	0.4269	0.4030	0.4293	0.4021	0.4287
%C. 3PTS		-5.6	0.6	-5.8	0.4
E. EX.	0.5137	0.4762	0.5062	0.4793	0.5155
P. EX.	0.2850	0.2614	0.2817	0.2626	0.2855

Cuadro B.24.: Batería II para la red simple de CRANFIELD: *pp2* y *fp10*, con y sin *qf*, con y sin diferencia de probabilidades.

EXP.	SMART	0, –	0	1, –	1
REL. REC.	260	242	253	235	246
EXHAUST.	PRECISIÓN				
0.00	0.9137	0.9060	0.9132	0.9078	0.9041
0.10	0.8181	0.8403	0.8194	0.8482	0.8181
0.20	0.7512	0.7022	0.7423	0.7019	0.7303
0.30	0.6858	0.6481	0.6968	0.6352	0.6770
0.40	0.6344	0.5383	0.6011	0.5350	0.5816
0.50	0.5543	0.4609	0.5039	0.4606	0.4942
0.60	0.4789	0.4116	0.4578	0.4087	0.4534
0.70	0.4297	0.3423	0.3946	0.3569	0.3994
0.80	0.3527	0.2614	0.3260	0.2686	0.3233
0.90	0.2527	0.1671	0.2221	0.1741	0.2190
1.00	0.1193	0.0872	0.1105	0.0878	0.1120
M. 11PTS	0.5446	0.4878	0.5262	0.4895	0.5193
%C. 11PTS		-10.4	-3.4	-10.1	-4.6
M. 3PTS	0.5527	0.4748	0.5241	0.4770	0.5159
%C. 3PTS		-14.1	-5.2	-13.7	-6.7
E. EX.	0.4104	0.3773	0.3931	0.3681	0.3829
P. EX.	0.5778	0.5378	0.5622	0.5222	0.5467

Cuadro B.25.: Batería II para la red simple de MEDLARS: *pp2* y *fp8*, con y sin *qf*, con y sin diferencia de probabilidades.

EXP.	SMART	0, –	0	1, –	1
REL. REC.	260	266	269	258	264
EXHAUST.	PRECISIÓN				
0.00	0.9137	0.9403	0.9259	0.9287	0.9139
0.10	0.8181	0.8559	0.8398	0.8415	0.8241
0.20	0.7512	0.7572	0.7584	0.7467	0.7543
0.30	0.6858	0.7057	0.7095	0.6930	0.6962
0.40	0.6344	0.6382	0.6621	0.6144	0.6384
0.50	0.5543	0.5453	0.5681	0.5209	0.5518
0.60	0.4789	0.4651	0.4996	0.4468	0.4784
0.70	0.4297	0.3761	0.4265	0.3832	0.4270
0.80	0.3527	0.2990	0.3530	0.3076	0.3507
0.90	0.2527	0.1996	0.2485	0.2140	0.2511
1.00	0.1193	0.0976	0.1157	0.1030	0.1177
M. 11PTS	0.5446	0.5345	0.5552	0.5272	0.5458
%C. 11PTS		-1.8	1.9	-3.2	0.2
M. 3PTS	0.5527	0.5338	0.5598	0.5250	0.5523
%C. 3PTS		-3.4	1.3	-5.0	-0.1
E. EX.	0.4104	0.4152	0.4212	0.4047	0.4156
P. EX.	0.5778	0.5911	0.5978	0.5733	0.5867

Cuadro B.26.: Batería II para la red simple de MEDLARS: *pp2* y *fp10*, con y sin *qf*, con y sin diferencia de probabilidades.

Exp.	SMART	*pc-mv*, 0	*pc-mv*, 1	*pc-J*, 0	*pc-J*, 1
REL. REC.	91	90	91	92	93
RECALL	PRECISIÓN				
0.00	0.7007	0.5101	0.6708	0.5967	0.6800
0.10	0.6911	0.4904	0.6512	0.5810	0.6619
0.20	0.6548	0.4691	0.6179	0.5628	0.6293
0.30	0.5806	0.4203	0.5487	0.4956	0.5693
0.40	0.5514	0.3839	0.5177	0.4668	0.5325
0.50	0.5177	0.3798	0.5163	0.4608	0.5250
0.60	0.4063	0.3180	0.4097	0.3874	0.4281
0.70	0.2979	0.2501	0.2942	0.2786	0.2948
0.80	0.2822	0.2391	0.2708	0.2602	0.2653
0.90	0.2486	0.1954	0.2313	0.2272	0.2447
1.00	0.2454	0.1948	0.2308	0.2259	0.2431
M. 11PTS	0.4706	0.3501	0.4509	0.4130	0.4613
%C. 11PTS		-25.6	-4.2	-12.2	-2.0
M. 3PTS	0.4849	0.3627	0.4683	0.4279	0.4732
%C. 3PTS		-25.2	-3.4	-11.8	-2.4
R. EX.	0.6120	0.6055	0.6161	0.5999	0.6196
P. EX.	0.2719	0.2314	0.2628	0.2615	0.2774

Cuadro B.27.: Batería I para la red aumentada de ADI: *pp2, fp10*, *pc-mv, pc-J*, con y sin *qf*.

Exp.	SMART	*pc-mv*, 0	*pc-mv*, 1	*pc-J*, 0	*pc-J*, 1
REL. REC.	246	223	247	230	244
RECALL	PRECISIÓN				
0.00	0.7375	0.7520	0.7562	0.7810	0.7512
0.10	0.6806	0.6692	0.6858	0.6926	0.6840
0.20	0.5742	0.5539	0.6067	0.5800	0.6078
0.30	0.4774	0.4853	0.5222	0.5070	0.5256
0.40	0.4168	0.3822	0.4685	0.3936	0.4652
0.50	0.3515	0.3235	0.3824	0.3352	0.3811
0.60	0.2808	0.2546	0.3105	0.2737	0.3107
0.70	0.2192	0.1886	0.2404	0.2028	0.2411
0.80	0.1793	0.1540	0.2048	0.1671	0.2062
0.90	0.1255	0.1000	0.1474	0.1134	0.1464
1.00	0.1020	0.0772	0.1197	0.0885	0.1201
M. 11PTS	0.3768	0.3582	0.4041	0.3759	0.4046
%C. 11PTS		-4.9	7.2	-0.2	7.3
M. 3PTS	0.3683	0.3438	0.3980	0.3608	0.3983
%C. 3PTS		-6.7	8.1	-2.0	8.2
R. EX.	0.4172	0.3733	0.4182	0.3793	0.4141
P. EX.	0.3726	0.3244	0.3892	0.3412	0.3854

Cuadro B.28.: Batería I para la red aumentada de CACM: *pp2, fp10*, *pc-mv, pc-J*, con y sin *qf*.

Exp.	SMART	*pc-mv*, 0	*pc-mv*, 1	*pc-J*, 0	*pc-J*, 1
REL. REC.	343	277	318	282	318
RECALL	PRECISIÓN				
0.00	0.6129	0.5925	0.5670	0.5906	0.5752
0.10	0.4485	0.3543	0.4101	0.3690	0.4172
0.20	0.3696	0.2754	0.3369	0.2901	0.3406
0.30	0.2950	0.2165	0.2727	0.2341	0.2748
0.40	0.2491	0.1777	0.2376	0.1911	0.2416
0.50	0.2120	0.1505	0.1983	0.1647	0.2047
0.60	0.1766	0.1242	0.1653	0.1259	0.1602
0.70	0.1285	0.0918	0.1228	0.0892	0.1151
0.80	0.0999	0.0708	0.0941	0.0691	0.0918
0.90	0.0689	0.0519	0.0716	0.0488	0.0689
1.00	0.0444	0.0369	0.0424	0.0357	0.0408
M. 11PTS	0.2459	0.1948	0.2290	0.2007	0.2301
%C. 11PTS		-20.8	-6.9	-18.4	-6.4
M. 3PTS	0.2272	0.1656	0.2098	0.1746	0.2124
%C. 3PTS		-27.1	-7.6	-23.1	-6.5
R. EX.	0.1810	0.1237	0.1531	0.1270	0.1503
P. EX.	0.3066	0.2430	0.2789	0.2474	0.2789

Cuadro B.29.: Batería I para la red aumentada de CISI: *pp2, fp10*, *pc-mv, pc-J*, con y sin *qf*.

Exp.	SMART	*pc-mv*, 0	*pc-mv*, 1	*pc-J*, 0	*pc-J*, 1
REL. REC.	824	812	814	826	809
RECALL	PRECISIÓN				
0.00	0.8020	0.7990	0.7676	0.8070	0.7633
0.10	0.7780	0.7689	0.7458	0.7822	0.7400
0.20	0.6651	0.6590	0.6396	0.6713	0.6353
0.30	0.5598	0.5472	0.5332	0.5591	0.5334
0.40	0.4656	0.4473	0.4442	0.4574	0.4413
0.50	0.4131	0.3948	0.3958	0.4060	0.3918
0.60	0.3195	0.3148	0.3189	0.3282	0.3184
0.70	0.2467	0.2438	0.2357	0.2518	0.2362
0.80	0.2024	0.1992	0.1981	0.2057	0.1970
0.90	0.1445	0.1418	0.1429	0.1459	0.1427
1.00	0.1267	0.1260	0.1282	0.1310	0.1280
M. 11PTS	0.4294	0.4220	0.4136	0.4314	0.4116
%C. 11PTS		-1.7	-3.7	0.5	-4.1
M. 3PTS	0.4269	0.4177	0.4112	0.4277	0.4081
%C. 3PTS		-2.1	-3.7	0.2	-4.4
R. EX.	0.5137	0.4999	0.5113	0.5087	0.5090
P. EX.	0.2850	0.2803	0.2762	0.2861	0.2749

Cuadro B.30.: Batería I para la red aumentada de CRANFIELD: *pp2, fp10*, *pc-mv, pc-J*, con y sin *qf*.

Exp.	SMART	*pc-mv*, 0	*pc-mv*, 1	*pc-J*, 0	*pc-J*, 1
REL. REC.	260	288	265	292	266
RECALL	PRECISIÓN				
0.00	0.9137	0.9554	0.8931	0.9587	0.8933
0.10	0.8181	0.9038	0.8156	0.8812	0.8109
0.20	0.7512	0.8150	0.7540	0.8158	0.7496
0.30	0.6858	0.7658	0.7229	0.7720	0.7215
0.40	0.6344	0.7089	0.6528	0.7114	0.6548
0.50	0.5543	0.6506	0.6009	0.6548	0.5941
0.60	0.4789	0.5855	0.5543	0.6028	0.5531
0.70	0.4297	0.4932	0.5107	0.5051	0.5044
0.80	0.3527	0.4256	0.4343	0.4446	0.4194
0.90	0.2527	0.2778	0.2997	0.3018	0.2967
1.00	0.1193	0.1658	0.1816	0.1718	0.1739
M. 11PTS	0.5446	0.6134	0.5836	0.6200	0.5792
%C. 11PTS		12.6	7.2	13.8	6.4
M. 3PTS	0.5527	0.6304	0.5964	0.6384	0.5877
%C. 3PTS		14.1	7.9	15.5	6.3
R. EX.	0.4104	0.4496	0.4212	0.4580	0.4229
P. EX.	0.5778	0.6400	0.5889	0.6489	0.5911

Cuadro B.31.: Batería I para la red aumentada de MEDLARS: *pp2, fp10*, *pc-mv, pc-J*, con y sin *qf*.

EXP.	SMART	fp10	fp10-
REL. REC.	91	93	91
EXHAUST.	PRECISIÓN		
0.00	0.7007	0.6800	0.6650
0.10	0.6911	0.6619	0.6457
0.20	0.6548	0.6293	0.6355
0.30	0.5806	0.5693	0.5738
0.40	0.5514	0.5325	0.5238
0.50	0.5177	0.5250	0.5167
0.60	0.4063	0.4281	0.4198
0.70	0.2979	0.2948	0.2971
0.80	0.2822	0.2653	0.2673
0.90	0.2486	0.2447	0.2477
1.00	0.2454	0.2431	0.2462
M. 11PTS	0.4706	0.4613	0.4581
%C. 11PTS		-2.0	-2.7
M. 3PTS	0.4849	0.4732	0.4732
%C. 3PTS		-2.5	-2.5
E. EX.	0.6120	0.6196	0.6053
P. EX.	0.1733	0.1771	0.1733

Cuadro B.32.: Batería II para la red aumentada de ADI: *pp2, pc-J, con replicación* y *fp10 vs fpl0-*.

EXP.	SMART	fp10	fp10-
REL. REC.	246	244	242
RECALL	PRECISIÓN		
0.00	0.7375	0.7512	0.7371
0.10	0.6806	0.6840	0.6722
0.20	0.5742	0.6078	0.6056
0.30	0.4774	0.5256	0.5197
0.40	0.4168	0.4652	0.4610
0.50	0.3515	0.3811	0.3763
0.60	0.2808	0.3107	0.3064
0.70	0.2192	0.2411	0.2409
0.80	0.1793	0.2062	0.2081
0.90	0.1255	0.1464	0.1478
1.00	0.1020	0.1201	0.1211
M. 11PTS	0.3768	0.4046	0.3996
%C. 11PTS		7.3	6.0
M. 3PTS	0.3683	0.3983	0.3967
%C. 3PTS		8.2	7.7
R. EX.	0.4172	0.4141	0.4121
P. EX.	0.3154	0.3854	0.3103

Cuadro B.33.: Batería II para la red aumentada de CACM: *pp2, pc-J, con replicación* y *fp10 vs fpl0-*.

EXP.	SMART	fp10	fp10-
REL. REC.	343	318	309
RECALL	PRECISIÓN		
0.00	0.6129	0.5752	0.5567
0.10	0.4485	0.4172	0.4122
0.20	0.3696	0.3406	0.3389
0.30	0.2950	0.2748	0.2757
0.40	0.2491	0.2416	0.2407
0.50	0.2120	0.2047	0.2063
0.60	0.1766	0.1602	0.1625
0.70	0.1285	0.1151	0.1207
0.80	0.0999	0.0918	0.0967
0.90	0.0689	0.0689	0.0737
1.00	0.0444	0.0408	0.0445
M. 11PTS	0.2459	0.2301	0.2299
%C. 11PTS		-6.4	-6.5
M. 3PTS	0.2272	0.2124	0.2140
%C. 3PTS		-6.5	-5.8
R. EX.	0.1810	0.1503	0.1496
P. EX.	0.3009	0.2789	0.2710

Cuadro B.34.: Batería II para la red aumentada de CISI: *pp2, pc-J, con replicación* y *fp10 vs fpl0-*.

EXP.	SMART	fp10	fp10-
REL. REC.	824	826	847
RECALL	PRECISIÓN		
0.00	0.8020	0.8070	0.8083
0.10	0.7780	0.7822	0.7881
0.20	0.6651	0.6713	0.6679
0.30	0.5598	0.5591	0.5585
0.40	0.4656	0.4574	0.4643
0.50	0.4131	0.4060	0.4216
0.60	0.3195	0.3282	0.3414
0.70	0.2467	0.2518	0.2733
0.80	0.2024	0.2057	0.2262
0.90	0.1445	0.1459	0.1646
1.00	0.1267	).1310	0.1484
M. 11PTS	0.4294	0.4315	0.4421
%C. 11PTS		0.5	2.9
M. 3PTS	0.4269	0.4277	0.4386
%C. 3PTS		0.2	2.7
R. EX.	0.5137	0.5087	0.5236
P. EX.	0.2441	0.2861	0.2510

Cuadro B.35.: Batería II para la red aumentada de CRANFIELD: *pp2, pc-J, sin replicación* y *fp10 vs fpl0-*.

EXP.	SMART	fp10	fp10-
REL. REC.	260	292	293
RECALL	PRECISIÓN		
0.00	0.9137	0.9587	0.9615
0.10	0.8181	0.8812	0.8811
0.20	0.7512	0.8158	0.8198
0.30	0.6858	0.7720	0.7814
0.40	0.6344	0.7114	0.7239
0.50	0.5543	0.6548	0.6775
0.60	0.4789	0.6028	0.6063
0.70	0.4297	0.5051	0.5477
0.80	0.3527	0.4446	0.4839
0.90	0.2527	0.3018	0.3553
1.00	0.1193	0.1718	0.2088
M. 11PTS	0.5446	0.6200	0.6407
%C. 11PTS		13.8	17.6
M. 3PTS	0.5527	0.6384	0.6604
%C. 3PTS		15.5	19.4
R. EX.	0.4104	0.4580	0.4587
P. EX.	0.5778	0.6489	0.6511

Cuadro B.36.: Batería II para la red aumentada de MEDLARS: *pp2, pc-J, sin replicación* y *fp10 vs fpl0-*.

Exp.	Smart	$\alpha = 0{,}2$	$\alpha = 0{,}5$	$\alpha = 0{,}8$
Rel. Rec.	91	90	91	91
Recall	Precisión			
0.00	0.7007	0.6652	0.6644	0.6668
0.10	0.6911	0.6470	0.6439	0.6471
0.20	0.6548	0.6324	0.6340	0.6369
0.30	0.5806	0.5781	0.5747	0.5752
0.40	0.5514	0.5244	0.5236	0.5252
0.50	0.5177	0.5223	0.5194	0.5185
0.60	0.4063	0.4275	0.4246	0.4248
0.70	0.2979	0.3076	0.2998	0.3013
0.80	0.2822	0.2782	0.2723	0.2723
0.90	0.2486	0.2553	0.2518	0.2509
1.00	0.2454	0.2530	0.2490	0.2490
M. 11pts	0.4706	0.4628	0.4598	0.4607
%C. 11pts		-1.7	-2.3	-2.2
M. 3pts	0.4849	0.4776	0.4753	0.4759
%C. 3pts		-1.6	-2.0	-1.9
R. Ex.	0.6120	0.6027	0.6053	0.6053
P. Ex.	0.1733	0.1714	0.1733	0.1733

Cuadro B.37.: Batería III para la red aumentada de ADI: *pp2, pc-J, fp10-, con replicación e instanciación parcial*.

Exp.	Smart	$\alpha = 0{,}2$	$\alpha = 0{,}5$	$\alpha = 0{,}8$
Rel. Rec.	246	220	228	236
Recall	Precisión			
0.00	0.7375	0.7059	0.7254	0.7451
0.10	0.6806	0.6441	0.6651	0.6822
0.20	0.5742	0.5608	0.5746	0.5989
0.30	0.4774	0.4792	0.4910	0.5097
0.40	0.4168	0.4302	0.4415	0.4571
0.50	0.3515	0.3678	0.3738	0.3798
0.60	0.2808	0.3088	0.3124	0.3115
0.70	0.2192	0.2424	0.2402	0.2409
0.80	0.1793	0.2116	0.2114	0.2099
0.90	0.1255	0.1513	0.1498	0.1486
1.00	0.1020	0.1196	0.1201	0.1213
M. 11pts	0.3768	0.3838	0.3914	0.4005
%C. 11pts		1.8	3.8	6.2
M. 3pts	0.3683	0.3800	0.3866	0.3962
%C. 3pts		3.1	4.9	7.5
R. Ex.	0.4172	0.3874	0.3963	0.4072
P. Ex.	0.3154	0.2821	0.2923	0.3026

Cuadro B.38.: Batería III para la red aumentada de CACM: *pp2, pc-J, fp10-, con replicación e instanciación parcial*.

EXP.	SMART	$\alpha = 0{,}2$	$\alpha = 0{,}5$	$\alpha = 0{,}8$
REL. REC.	343	308	287	268
RECALL	PRECISIÓN			
0.00	0.6129	0.5238	0.4890	0.4683
0.10	0.4485	0.3843	0.3500	0.3296
0.20	0.3696	0.3304	0.2882	0.2700
0.30	0.2950	0.2723	0.2397	0.2258
0.40	0.2491	0.2423	0.2149	0.1972
0.50	0.2120	0.2135	0.1828	0.1667
0.60	0.1766	0.1832	0.1642	0.1476
0.70	0.1285	0.1367	0.1214	0.1090
0.80	0.0999	0.1047	0.0936	0.0876
0.90	0.0689	0.0812	0.0715	0.0674
1.00	0.0444	0.0474	0.0399	0.0399
M. 11PTS	0.2459	0.2291	0.2050	0.1917
%C. 11PTS		-6.9	-16.7	-22.1
M. 3PTS	0.2272	0.2162	0.1882	0.1748
%C. 3PTS		-4.9	-17.2	-23.1
R. EX.	0.1810	0.1597	0.1310	0.1205
P. EX.	0.3009	0.2702	0.2518	0.2351

Cuadro B.39.: Batería III para la red aumentada de CISI: *pp2, pc-J, fp10-, con replicación e instanciación parcial*.

EXP.	SMART	$\alpha = 0{,}2$	$\alpha = 0{,}5$	$\alpha = 0{,}8$
REL. REC.	824	819	839	842
RECALL	PRECISIÓN			
0.00	0.8020	0.7767	0.7948	0.8025
0.10	0.7780	0.7498	0.7703	0.7784
0.20	0.6651	0.6426	0.6540	0.6656
0.30	0.5598	0.5392	0.5494	0.5545
0.40	0.4656	0.4488	0.4571	0.4599
0.50	0.4131	0.4044	0.4152	0.4189
0.60	0.3195	0.3336	0.3391	0.3412
0.70	0.2467	0.2603	0.2665	0.2703
0.80	0.2024	0.2140	0.2206	0.2239
0.90	0.1445	0.1592	0.1637	0.1658
1.00	0.1267	0.1438	0.1480	0.1494
M. 11PTS	0.4294	0.4248	0.4344	0.4391
%C. 11PTS		-1.1	1.1	2.2
M. 3PTS	0.4269	0.4203	0.4299	0.4361
%C. 3PTS		-1.6	.7	2.1
R. EX.	0.5137	0.5109	0.5225	0.5213
P. EX.	0.2441	0.2427	0.2486	0.2495

Cuadro B.40.: Batería III para la red aumentada de CRANFIELD: *pp2, pc-J, fp10-, sin replicación e instanciación parcial*.

EXP.	SMART	$\alpha = 0{,}2$	$\alpha = 0{,}5$	$\alpha = 0{,}8$
REL. REC.	260	287	287	293
RECALL	PRECISIÓN			
0.00	0.9137	0.9451	0.9590	0.9597
0.10	0.8181	0.8596	0.8732	0.8748
0.20	0.7512	0.8192	0.8179	0.8221
0.30	0.6858	0.7667	0.7721	0.7770
0.40	0.6344	0.7242	0.7278	0.7263
0.50	0.5543	0.6810	0.6819	0.6805
0.60	0.4789	0.6152	0.6162	0.6092
0.70	0.4297	0.5546	0.5560	0.5547
0.80	0.3527	0.5097	0.4999	0.4932
0.90	0.2527	0.3750	0.3686	0.3608
1.00	0.1193	0.2202	0.2145	0.2105
M. 11PTS	0.5446	0.6428	0.6443	0.6426
%C. 11PTS		18.0	18.3	17.9
M. 3PTS	0.5527	0.6700	0.6666	0.6653
%C. 3PTS		21.2	20.6	20.3
R. EX.	0.4104	0.4514	0.4503	0.4587
P. EX.	0.5778	0.6378	0.6378	0.6511

Cuadro B.41.: Batería III para la red aumentada de MEDLARS: *pp2, pc-J, fp10-, sin replicación e instanciación parcial*.

EXP.	SMART	fp10-	fp10	fp10c-	fp10c	fp10d-	fp10d	fp10e-	fp10e
REL. REC.	91	88	88	73	68	84	82	88	88
EXHAUST.	PRECISIÓN								
0.00	0.7007	0.6836	0.6859	0.6017	0.5767	0.7113	0.7095	0.6836	0.6858
0.10	0.6911	0.6686	0.6693	0.5888	0.5767	0.6947	0.6917	0.6686	0.6694
0.20	0.6548	0.6422	0.6426	0.5425	0.5338	0.6354	0.6305	0.6422	0.6427
0.30	0.5806	0.5757	0.5773	0.4806	0.4741	0.5610	0.5601	0.5757	0.5774
0.40	0.5514	0.5146	0.5183	0.4705	0.4551	0.5234	0.5207	0.5146	0.5178
0.50	0.5177	0.5025	0.5031	0.4656	0.4497	0.5121	0.5109	0.5025	0.5036
0.60	0.4063	0.3772	0.3752	0.3437	0.3489	0.3878	0.3864	0.3772	0.3762
0.70	0.2979	0.2737	0.2730	0.2163	0.2180	0.2663	0.2681	0.2737	0.2741
0.80	0.2822	0.2460	0.2428	0.2142	0.2049	0.2358	0.2320	0.2460	0.2439
0.90	0.2486	0.2175	0.2092	0.1878	0.1781	0.2125	0.2076	0.2175	0.2097
1.00	0.2454	0.2156	0.2088	0.1863	0.1781	0.2110	0.2076	0.2156	0.2093
M. 11PTS	0.4706	0.4470	0.4459	0.3907	0.3813	0.4501	0.4477	0.4470	0.4464
%C. 11PTS		-5.1	-5.3	-17.0	-19.0	-4.4	-4.9	-5.1	-5.2
M. 3PTS	0.4849	0.4636	0.4628	0.4075	0.3961	0.4611	0.4578	0.4636	0.4634
%C. 3PTS		-4.4	-4.6	-16.0	-18.4	-5.0	-5.6	-4.4	-4.5
R. EX.	0.6120	0.5875	0.5936	0.4553	0.4590	0.5720	0.5672	0.5875	0.5936
P. EX.	0.1733	0.1676	0.1676	0.1390	0.1295	0.1600	0.1562	0.1676	0.1676

Cuadro B.42.: Batería I para la red mixta de ADI: *pp2, pc-J, con replicación*. Comparación de las funciones de probabilidad *fp10, fp10c, fp10e, fp10e* y sus variaciones con las diferencias.

EXP.	SMART	fp10-	fp10	fp10c-	fp10c	fp10d-	fp10d	fp10e-	fp10e
REL. REC.	246	240	241	197	199	219	224	240	241
EXHAUST.	PRECISIÓN								
0.00	0.7375	0.7475	0.7458	0.4985	0.5071	0.7086	0.7138	0.7475	0.7494
0.10	0.6806	0.6871	0.6803	0.4447	0.4569	0.6351	0.6452	0.6871	0.6861
0.20	0.5742	0.6054	0.6015	0.3738	0.3788	0.5435	0.5476	0.6054	0.6053
0.30	0.4774	0.5021	0.5151	0.3276	0.3386	0.4650	0.4721	0.5021	0.5182
0.40	0.4168	0.4447	0.4573	0.2840	0.2866	0.4132	0.4145	0.4447	0.4575
0.50	0.3515	0.3697	0.3723	0.2425	0.2448	0.3473	0.3518	0.3695	0.3730
0.60	0.2808	0.3015	0.3028	0.1844	0.1850	0.2772	0.2809	0.3015	0.3038
0.70	0.2192	0.2325	0.2348	0.1387	0.1403	0.2170	0.2204	0.2325	0.2339
0.80	0.1793	0.2016	0.2015	0.1033	0.1041	0.1835	0.1859	0.2016	0.2013
0.90	0.1255	0.1436	0.1419	0.0829	0.0768	0.1378	0.1396	0.1436	0.1422
1.00	0.1020	0.1162	0.1161	0.0564	0.0501	0.1055	0.1075	0.1161	0.1164
M. 11PTS	0.3768	0.3956	0.3972	0.2488	0.2517	0.3667	0.3708	0.3956	0.3988
%C. 11PTS		4.9	5.4	-34.0	-33.3	-2.7	-1.6	4.9	5.8
M. 3PTS	0.3683	0.3922	0.3918	0.2399	0.2426	0.3581	0.3618	0.3922	0.3932
%C. 3PTS		6.4	6.3	-34.9	-34.2	-2.8	-1.8	6.4	6.7
R. EX.	0.4172	0.4086	0.4103	0.2813	0.2896	0.3872	0.3957	0.4086	0.4103
P. EX.	0.3154	0.3077	0.3090	0.2526	0.2551	0.2808	0.2872	0.3077	0.3090

Cuadro B.43.: Batería I para la red mixta de CACM: *pp2, pc-J, con replicación*. Comparación de las funciones de probabilidad *fp10, fp10c, fp10e, fp10e* y sus variaciones con las diferencias.

EXP.	SMART	fp10-	fp10	fp10c-	fp10c	fp10d-	fp10d	fp10e-	fp10e
REL. REC.	343	340	344	293	299	332	336	340	344
EXHAUST.	PRECISIÓN								
0.00	0.6129	0.5908	0.6003	0.5167	0.5212	0.6043	0.6078	0.5908	0.6003
0.10	0.4485	0.4602	0.4562	0.3648	0.3706	0.4521	0.4601	0.4602	0.4562
0.20	0.3696	0.3784	0.3783	0.3151	0.3204	0.3703	0.3740	0.3788	0.3782
0.30	0.2950	0.2978	0.2993	0.2620	0.2657	0.3027	0.3028	0.2978	0.2991
0.40	0.2491	0.2607	0.2593	0.2223	0.2226	0.2606	0.2603	0.2607	0.2594
0.50	0.2120	0.2242	0.2210	0.1848	0.1842	0.2243	0.2231	0.2242	0.2210
0.60	0.1766	0.1863	0.1824	0.1508	0.1487	0.1827	0.1806	0.1863	0.1825
0.70	0.1285	0.1398	0.1346	0.1137	0.1104	0.1442	0.1396	0.1398	0.1346
0.80	0.0999	0.1063	0.1027	0.0895	0.0848	0.1122	0.1097	0.1063	0.1027
0.90	0.0689	0.0792	0.0747	0.0610	0.0568	0.0793	0.0755	0.0792	0.0747
1.00	0.0444	0.0494	0.0475	0.0372	0.0366	0.0491	0.0477	0.0494	0.0475
M. 11PTS	0.2459	0.2521	0.2506	0.2107	0.2111	0.2529	0.2528	0.2521	0.2506
%C. 11PTS		2.5	1.9	-14.4	-14.2	2.8	2.8	2.5	1.9
M. 3PTS	0.2272	0.2363	0.2340	0.1965	0.1965	0.2356	0.2356	0.2365	0.2340
%C. 3PTS		4.0	2.9	-13.6	-13.6	3.6	3.6	4.0	2.9
R. EX.	0.1810	0.1762	0.1774	0.1386	0.1390	0.1741	0.1760	0.1762	0.1774
P. EX.	0.3009	0.2982	0.3018	0.2570	0.2623	0.2912	0.2947	0.2982	0.3018

Cuadro B.44.: Batería I para la red mixta de CISI: *pp2, pc-J, con replicación*. Comparación de las funciones de probabilidad *fp10, fp10c, fp10e, fp10e* y sus variaciones con las diferencias.

EXP.	SMART	fp10-	fp10	fp10c-	fp10c	fp10d-	fp10d	fp10e-	fp10e
REL. REC.	824	831	813	704	717	836	829	829	809
EXHAUST.	PRECISIÓN								
0.00	0.8020	0.7960	0.8074	0.7027	0.7030	0.8113	0.8071	0.7960	0.8072
0.10	0.7780	0.7774	0.7836	0.6657	0.6717	0.7786	0.7826	0.7759	0.7823
0.20	0.6651	0.6623	0.6631	0.5509	0.5494	0.6698	0.6743	0.6609	0.6597
0.30	0.5598	0.5587	0.5531	0.4516	0.4460	0.5649	0.5644	0.5585	0.5549
0.40	0.4656	0.4712	0.4531	0.3602	0.3604	0.4784	0.4801	0.4707	0.4544
0.50	0.4131	0.4228	0.4060	0.3186	0.3167	0.4315	0.4249	0.4222	0.4079
0.60	0.3195	0.3353	0.3174	0.2632	0.2627	0.3490	0.3403	0.3346	0.3186
0.70	0.2467	0.2554	0.2403	0.1986	0.1968	0.2722	0.2599	0.2548	0.2401
0.80	0.2024	0.2122	0.1981	0.1675	0.1631	0.2193	0.2097	0.2119	0.1980
0.90	0.1445	0.1532	0.1368	0.1235	0.1187	0.1591	0.1502	0.1532	0.1372
1.00	0.1267	0.1371	0.1235	0.1094	0.1039	0.1431	0.1349	0.1370	0.1237
M. 11PTS	0.4294	0.4347	0.4257	0.3556	0.3539	0.4434	0.4389	0.4342	0.4258
%C. 11PTS		1.2	-.9	-17.2	-17.6	3.2	2.2	1.1	-.9
M. 3PTS	0.4269	0.4324	0.4224	0.3456	0.3431	0.4402	0.4363	0.4317	0.4219
%C. 3PTS		1.2	-1.1	-19.1	-19.7	3.1	2.2	1.1	-1.2
R. EX.	0.5137	0.5136	0.5002	0.4379	0.4435	0.5213	0.5150	0.5128	0.4985
P. EX.	0.2441	0.2462	0.2409	0.2086	0.2124	0.2477	0.2456	0.2456	0.2397

Cuadro B.45.: Batería I para la red mixta de CRANFIELD: *pp2, pc-J, sin replicación*. Comparación de las funciones de probabilidad *fp10, fp10c, fp10e, fp10e* y sus variaciones con las diferencias.

EXP.	SMART	fp10-	fp10	fp10c-	fp10c	fp10d-	fp10d	fp10e-	fp10e
REL. REC.	260	285	277	288	288	287	287	285	277
EXHAUST.	PRECISIÓN								
0.00	0.9137	0.9389	0.9406	0.9522	0.9566	0.9450	0.9493	0.9389	0.9406
0.10	0.8181	0.8594	0.8510	0.8753	0.8977	0.8592	0.8826	0.8594	0.8510
0.20	0.7512	0.8064	0.7977	0.8279	0.8237	0.8218	0.8127	0.8064	0.7977
0.30	0.6858	0.7609	0.7514	0.7590	0.7576	0.7697	0.7612	0.7609	0.7514
0.40	0.6344	0.6881	0.6823	0.6983	0.6987	0.7203	0.7165	0.6881	0.6820
0.50	0.5543	0.6313	0.6165	0.6465	0.6411	0.6547	0.6500	0.6313	0.6162
0.60	0.4789	0.5742	0.5581	0.6020	0.5830	0.5927	0.5865	0.5742	0.5582
0.70	0.4297	0.5274	0.4930	0.5003	0.4993	0.5512	0.5410	0.5274	0.4932
0.80	0.3527	0.4351	0.3928	0.4181	0.4023	0.4687	0.4384	0.4351	0.3927
0.90	0.2527	0.3196	0.2676	0.3056	0.2722	0.3364	0.3012	0.3196	0.2675
1.00	0.1193	0.1775	0.1512	0.1697	0.1505	0.1896	0.1674	0.1775	0.1511
M. 11PTS	0.5446	0.6108	0.5911	0.6141	0.6075	0.6281	0.6188	0.6108	0.5910
%C. 11PTS		12.1	8.5	12.7	11.5	15.3	13.6	12.1	8.5
M. 3PTS	0.5527	0.6243	0.6023	0.6308	0.6224	0.6484	0.6337	0.6243	0.6022
%C. 3PTS		12.9	8.9	14.1	12.6	17.3	14.6	12.9	8.9
R. EX.	0.4104	0.4452	0.4355	0.4509	0.4505	0.4538	0.4534	0.4452	0.4355
P. EX.	0.5778	0.6333	0.6156	0.6400	0.6400	0.6378	0.6378	0.6333	0.6156

Cuadro B.46.: Batería I para la red mixta de MEDLARS: *pp2, pc-J, sin replicación*. Comparación de las funciones de probabilidad *fp10, fp10c, fp10e, fp10e* y sus variaciones con las diferencias.

	SMART	$\alpha = 0{,}2$	$\alpha = 0{,}5$	$\alpha = 0{,}8$
REL. REC.	91	82	83	83
EXHAUST.	PRECISIÓN			
0.00	0.7007	0.6855	0.7001	0.7041
0.10	0.6911	0.6710	0.6835	0.6896
0.20	0.6548	0.6339	0.6468	0.6331
0.30	0.5806	0.5648	0.5790	0.5652
0.40	0.5514	0.5110	0.5262	0.5147
0.50	0.5177	0.5072	0.5210	0.5087
0.60	0.4063	0.3824	0.3791	0.3797
0.70	0.2979	0.2632	0.2649	0.2664
0.80	0.2822	0.2324	0.2329	0.2358
0.90	0.2486	0.2101	0.2107	0.2127
1.00	0.2454	0.2087	0.2090	0.2110
M. 11PTS	0.4706	0.4428	0.4503	0.4474
%C. 11PTS		-6.0	-4.4	-5.0
M. 3PTS	0.4849	0.4578	0.4669	0.4592
%C. 3PTS		-5.6	-3.8	-5.4
R. EX.	0.6120	0.5601	0.5648	0.5672
P. EX.	0.1733	0.1562	0.1581	0.1581

Cuadro B.47.: Batería II para la red mixta de ADI: *pp2, pc-J, con replicación, fp10d- e instanciación parcial*.

	SMART	$\alpha = 0,2$	$\alpha = 0,5$	$\alpha = 0,8$
REL. REC.	246	220	232	238
EXHAUST.	PRECISIÓN			
0.00	0.7375	0.6973	0.7258	0.7457
0.10	0.6806	0.6418	0.6598	0.6810
0.20	0.5742	0.5689	0.5836	0.6046
0.30	0.4774	0.4822	0.4961	0.5153
0.40	0.4168	0.4323	0.4430	0.4566
0.50	0.3515	0.3578	0.3706	0.3762
0.60	0.2808	0.3035	0.3088	0.3081
0.70	0.2192	0.2357	0.2328	0.2330
0.80	0.1793	0.2048	0.2033	0.2025
0.90	0.1255	0.1524	0.1496	0.1432
1.00	0.1020	0.1223	0.1211	0.1166
M. 11PTS	0.3768	0.3817	0.3904	0.3984
%C. 11PTS		1.3	3.6	5.7
M. 3PTS	0.3683	0.3772	0.3858	0.3944
%C. 3PTS		2.4	4.7	7.0
R. EX.	0.4172	0.3817	0.3975	0.4071
P. EX.	0.3154	0.2821	0.2974	0.3051

Cuadro B.48.: Batería II para la red mixta de CACM: *pp2, pc-J, con replicación, fp10e e instanciación parcial*.

	SMART	$\alpha = 0,2$	$\alpha = 0,5$	$\alpha = 0,8$
REL. REC.	343	339	333	308
EXHAUST.	PRECISIÓN			
0.00	0.6129	0.5887	0.5934	0.5809
0.10	0.4485	0.4459	0.4466	0.4155
0.20	0.3696	0.3753	0.3726	0.3452
0.30	0.2950	0.3038	0.3017	0.2791
0.40	0.2491	0.2705	0.2661	0.2410
0.50	0.2120	0.2313	0.2308	0.2095
0.60	0.1766	0.1963	0.1902	0.1762
0.70	0.1285	0.1462	0.1444	0.1361
0.80	0.0999	0.1159	0.1143	0.1084
0.90	0.0689	0.0822	0.0817	0.0786
1.00	0.0444	0.0498	0.0493	0.0484
M. 11PTS	0.2459	0.2551	0.2537	0.2381
%C. 11PTS		3.7	3.1	-3.2
M. 3PTS	0.2272	0.2408	0.2393	0.2210
%C. 3PTS		5.9	5.3	-2.8
R. EX.	0.1810	0.1816	0.1769	0.1649
P. EX.	0.3009	0.2974	0.2921	0.2702

Cuadro B.49.: Batería II para la red mixta de CISI: *pp2, pc-J, con replicación, fp10d- e instanciación parcial*.

	SMART	$\alpha = 0{,}2$	$\alpha = 0{,}5$	$\alpha = 0{,}8$
REL. REC.	824	803	833	840
EXHAUST.	PRECISIÓN			
0.00	0.8020	0.7727	0.7897	0.8025
0.10	0.7780	0.7395	0.7593	0.7685
0.20	0.6651	0.6269	0.6469	0.6606
0.30	0.5598	0.5314	0.5516	0.5627
0.40	0.4656	0.4445	0.4606	0.4766
0.50	0.4131	0.3987	0.4148	0.4307
0.60	0.3195	0.3212	0.3304	0.3440
0.70	0.2467	0.2477	0.2584	0.2663
0.80	0.2024	0.2007	0.2126	0.2162
0.90	0.1445	0.1456	0.1553	0.1570
1.00	0.1267	0.1301	0.1393	0.1410
M. 11PTS	0.4294	0.4144	0.4290	0.4387
%C. 11PTS		-3.5	-.1	2.1
M. 3PTS	0.4269	0.4087	0.4248	0.4358
%C. 3PTS		-4.3	-.5	2.0
R. EX.	0.5137	0.5049	0.5216	0.5234
P. EX.	0.2441	0.2379	0.2468	0.2489

Cuadro B.50.: Batería II para la red mixta de CRANFIELD: *pp2, pc-J, sin replicación, fp10d- e instanciación parcial*.

	SMART	$\alpha = 0{,}2$	$\alpha = 0{,}5$	$\alpha = 0{,}8$
REL. REC.	260	287	289	286
EXHAUST.	PRECISIÓN			
0.00	0.9137	0.9273	0.9459	0.9434
0.10	0.8181	0.8352	0.8427	0.8403
0.20	0.7512	0.7871	0.8162	0.8189
0.30	0.6858	0.7635	0.7691	0.7673
0.40	0.6344	0.7073	0.7131	0.7184
0.50	0.5543	0.6483	0.6538	0.6569
0.60	0.4789	0.6021	0.6018	0.5949
0.70	0.4297	0.5653	0.5605	0.5525
0.80	0.3527	0.4868	0.4792	0.4711
0.90	0.2527	0.3433	0.3471	0.3416
1.00	0.1193	0.2043	0.1989	0.1936
M. 11PTS	0.5446	0.6246	0.6298	0.6272
%C. 11PTS		14.6	15.6	15.1
M. 3PTS	0.5527	0.6408	0.6497	0.6490
%C. 3PTS		15.9	17.5	17.4
R. EX.	0.4104	0.4552	0.4599	0.4545
P. EX.	0.5778	0.6378	0.6422	0.6356

Cuadro B.51.: Batería II para la red mixta de MEDLARS: *pp2, pc-J, sin replicación, fp10d- e instanciación parcial*.

B.1. Curvas Exhaustividad - Precisión de los mejores resultados.

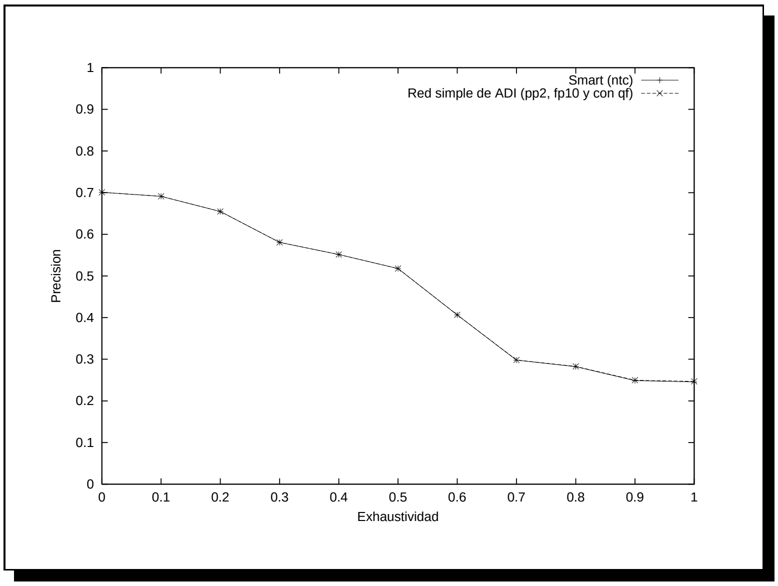


Figura B.1.: Curva Exhaustividad - Precisión para la mejor red en la colección ADI.

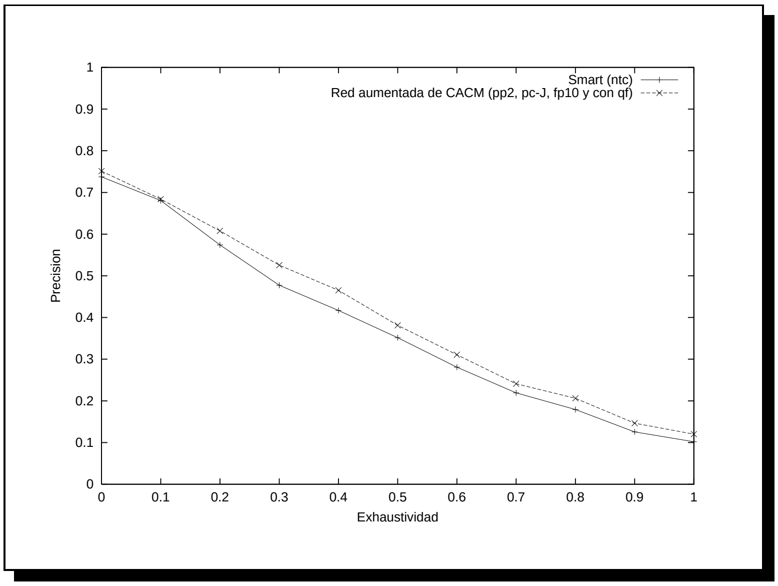


Figura B.2.: Curva Exhaustividad - Precisión para la mejor red en la colección CACM.

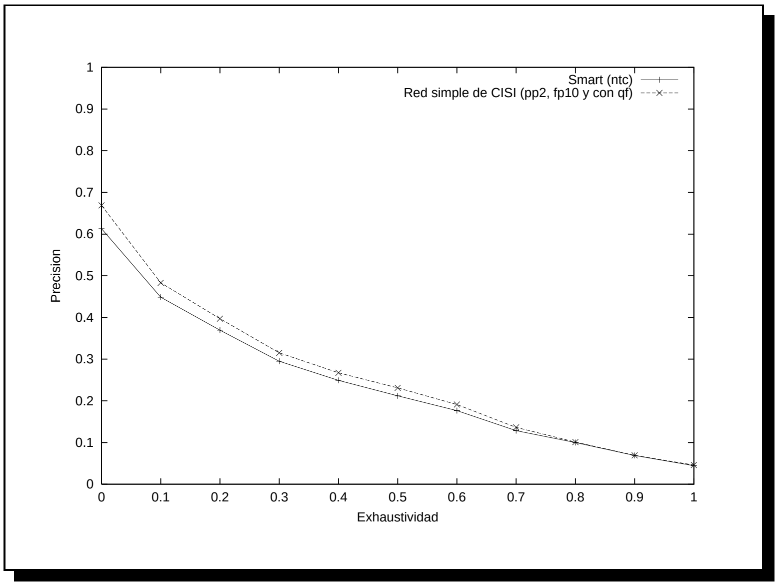


Figura B.3.: Curva Exhaustividad - Precisión para la mejor red en la colección CISI.

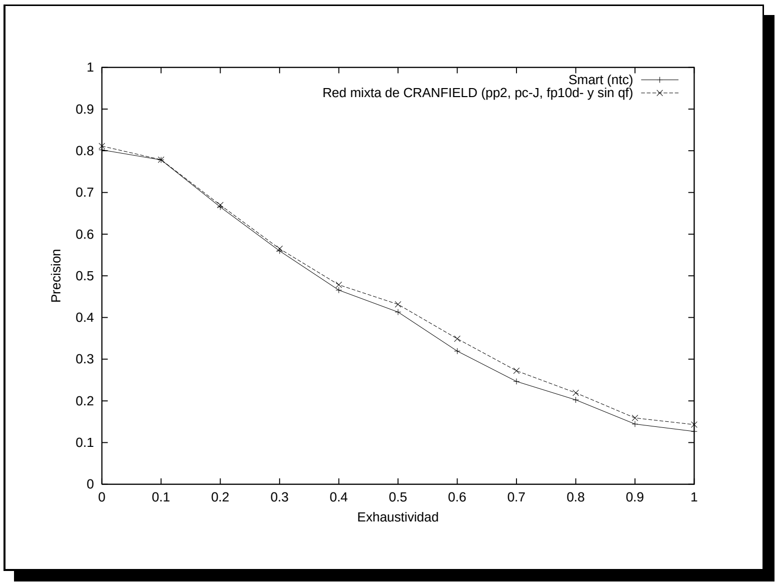


Figura B.4.: Curva Exhaustividad - Precisión para la mejor red en la colección CRANFIELD.

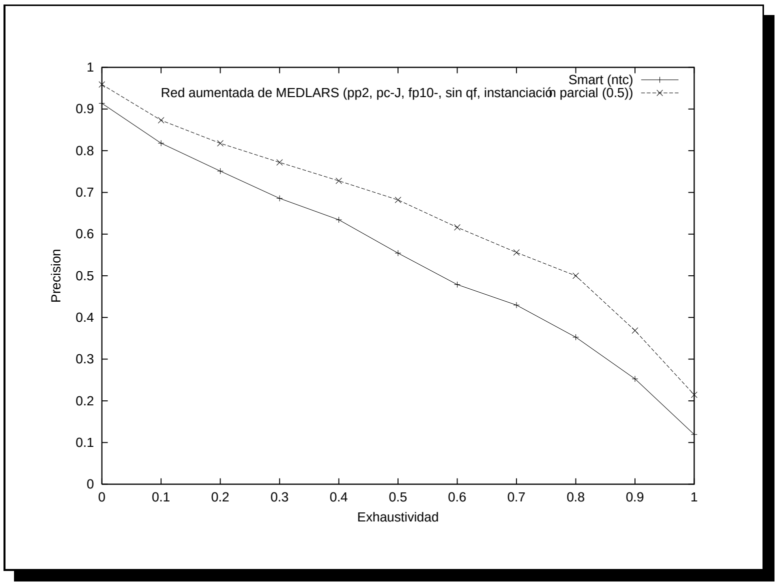


Figura B.5.: Curva Exhaustividad - Precisión para la mejor red en la colección MEDLARS.

C. Resultados empíricos detallados con la subred extendida de documentos.

EXP.	SMART	Mejor res.	$ndr = 5$	$ndr = 10$	$ndr = 15$
REL. REC.	91	91	92	96	94
RECALL	PRECISIÓN				
0.00	0.7007	0.7007	0.7055	0.7261	0.7142
0.10	0.6911	0.6911	0.6998	0.7180	0.7070
0.20	0.6548	0.6548	0.6831	0.6937	0.6832
0.30	0.5806	0.5806	0.6203	0.6358	0.6351
0.40	0.5514	0.5514	0.5877	0.5975	0.5972
0.50	0.5177	0.5177	0.5535	0.5679	0.5630
0.60	0.4063	0.4065	0.4549	0.4476	0.4324
0.70	0.2979	0.2982	0.3339	0.3210	0.3183
0.80	0.2822	0.2827	0.3266	0.3184	0.3128
0.90	0.2486	0.2497	0.2831	0.2913	0.2894
1.00	0.2454	0.2469	0.2764	0.2869	0.2855
M. 11PTS	0.4709	0.4706	0.5023	0.5095	0.5035
%C. 11PTS		0.1	6.7	8.2	6.9
M. 3PTS	0.4849	0.4851	0.5211	0.5267	0.5197
%C. 3PTS		0.0	7.4	8.6	7.1
R. EX.	0.6120	0.6120	0.6346	0.6131	0.6196
P. EX.	0.1733	0.1733	0.1829	0.1752	0.1790

Cuadro C.1.: Evaluación de la extensión de la subred de documentos para ADI.

EXP.	SMART	Mejor res.	$ndr = 5$	$ndr = 10$	$ndr = 15$
REL. REC.	246	244	222	233	236
RECALL	PRECISIÓN				
0.00	0.7375	0.7512	0.6855	0.6685	0.6598
0.10	0.6806	0.6840	0.6370	0.6070	0.6239
0.20	0.5742	0.6078	0.5290	0.5240	0.5147
0.30	0.4774	0.5256	0.4534	0.4590	0.4607
0.40	0.4168	0.4652	0.4101	0.4195	0.4089
0.50	0.3515	0.3811	0.3555	0.3600	0.3571
0.60	0.2808	0.3107	0.2858	0.2831	0.2825
0.70	0.2192	0.2411	0.2080	0.2126	0.2127
0.80	0.1793	0.2062	0.1838	0.1808	0.1694
0.90	0.1255	0.1464	0.1197	0.1215	0.1225
1.00	0.1020	0.1201	0.0942	0.0928	0.0899
M. 11PTS	0.3768	0.4036	0.3602	0.3572	0.3547
%C. 11PTS		7.1	-4.5	-5.3	-5.9
M. 3PTS	0.3683	0.3983	0.3561	0.3549	0.3470
%C. 3PTS		8.2	-3.4	-3.7	-5.8
R. EX.	0.4172	0.4141	0.3751	0.3869	0.3717
P. EX.	0.3154	0.3128	0.2846	0.2987	0.3026

Cuadro C.2.: Evaluación de la extensión de la subred de documentos para CACM.

EXP.	SMART	Mejor res.	$ndr = 5$	$ndr = 10$	$ndr = 15$
REL. REC.	343	369	363	370	376
RECALL	PRECISIÓN				
0.00	0.6129	0.6689	0.6408	0.6429	0.6348
0.10	0.4485	0.4831	0.4726	0.4801	0.4663
0.20	0.3696	0.3971	0.3927	0.4099	0.4082
0.30	0.2950	0.3151	0.3404	0.3477	0.3513
0.40	0.2491	0.2674	0.2761	0.2886	0.2895
0.50	0.2120	0.2311	0.2364	0.2579	0.2580
0.60	0.1766	0.1911	0.1937	0.2160	0.2168
0.70	0.1285	0.1364	0.1411	0.1646	0.1607
0.80	0.0999	0.1010	0.1103	0.1244	0.1185
0.90	0.0689	0.0692	0.0721	0.0884	0.0837
1.00	0.0444	0.0461	0.0504	0.0619	0.0577
M. 11PTS	0.2459	0.2642	0.2661	0.2802	0.2769
%C. 11PTS		7.4	8.2	13.9	12.6
M. 3PTS	0.2272	0.2430	0.2465	0.2641	0.2616
%C. 3PTS		7.0	8.4	16.2	15.1
R. EX.	0.1810	0.1881	0.1847	0.1893	0.2053
P. EX.	0.3009	0.3237	0.3184	0.3246	0.3298

Cuadro C.3.: Evaluación de la extensión de la subred de documentos para CISI.

EXP.	SMART	Mejor res.	$ndr = 5$	$ndr = 10$	$ndr = 15$
REL. REC.	824	836	543	576	574
RECALL	PRECISIÓN				
0.00	0.8020	0.8113	0.4235	0.4156	0.4264
0.10	0.7780	0.7786	0.3924	0.3788	0.3966
0.20	0.6651	0.6698	0.3339	0.3247	0.3354
0.30	0.5598	0.5649	0.2659	0.2731	0.2897
0.40	0.4656	0.4784	0.2323	0.2511	0.2690
0.50	0.4131	0.4315	0.2096	0.2356	0.2432
0.60	0.3195	0.3490	0.1814	0.2022	0.2092
0.70	0.2467	0.2722	0.1294	0.1595	0.1724
0.80	0.2024	0.2193	0.1044	0.1325	0.1454
0.90	0.1445	0.1591	0.0677	0.0916	0.1090
1.00	0.1267	0.1431	0.0602	0.0846	0.0986
M. 11PTS	0.4294	0.4434	0.2183	0.2318	0.2450
%C. 11PTS		3.2	-49.2	-46.1	-43.0
M. 3PTS	0.4269	0.4402	0.2160	0.2309	0.2413
%C. 3PTS		3.1	-49.5	-46.0	-43.5
R. EX.	0.5137	0.5213	0.3271	0.3485	0.3464
P. EX.	0.2441	0.2477	0.1609	0.1707	0.1701

Cuadro C.4.: Evaluación de la extensión de la subred de documentos para CRANFIELD.

EXP.	SMART	Mejor res.	$ndr = 5$	$ndr = 10$	$ndr = 15$
REL. REC.	260	287	266	279	269
RECALL	PRECISIÓN				
0.00	0.9137	0.9590	0.8123	0.8477	0.7553
0.10	0.8181	0.8732	0.7873	0.7930	0.7454
0.20	0.7512	0.8179	0.7617	0.7622	0.7137
0.30	0.6858	0.7721	0.7206	0.7197	0.6904
0.40	0.6344	0.7278	0.6713	0.6776	0.6734
0.50	0.5543	0.6819	0.6222	0.6511	0.6454
0.60	0.4789	0.6162	0.5560	0.6085	0.6134
0.70	0.4297	0.5560	0.5149	0.5482	0.5340
0.80	0.3527	0.4999	0.4059	0.4291	0.3774
0.90	0.2527	0.3686	0.2391	0.2565	0.2604
1.00	0.1193	0.2145	0.1023	0.0848	0.1143
M. 11PTS	0.5446	0.6443	0.5631	0.5799	0.5566
%C. 11PTS		18.3	3.3	6.4	2.2
M. 3PTS	0.5527	0.6666	0.5966	0.6142	0.5788
%C. 3PTS		20.6	7.9	11.1	4.7
R. EX.	0.4104	0.4503	0.4135	0.4476	0.4256
P. EX.	0.5778	0.6378	0.5911	0.6200	0.5978

Cuadro C.5.: Evaluación de la extensión de la subred de documentos para MEDLARS.

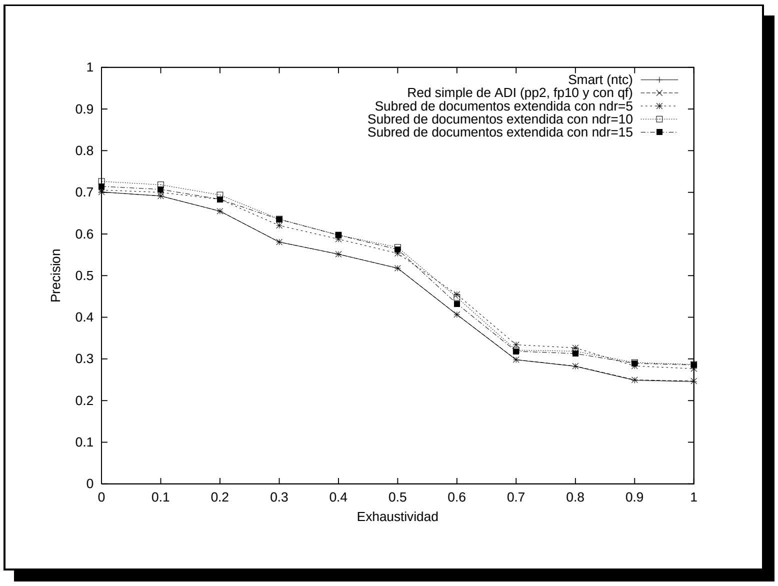


Figura C.1.: Curva Exhaustividad - Precisión para la subred extendida de documentos en la colección ADI.

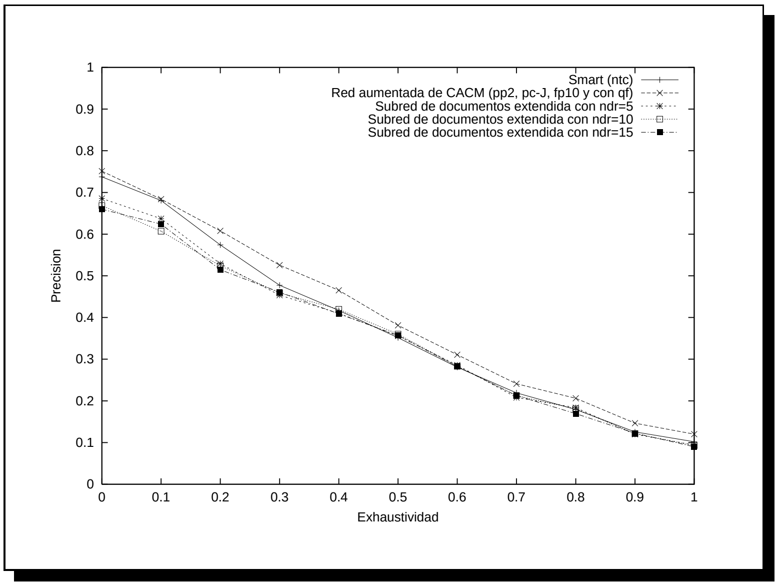


Figura C.2.: Curva Exhaustividad - Precisión para la subred extendida de documentos en la colección CACM.

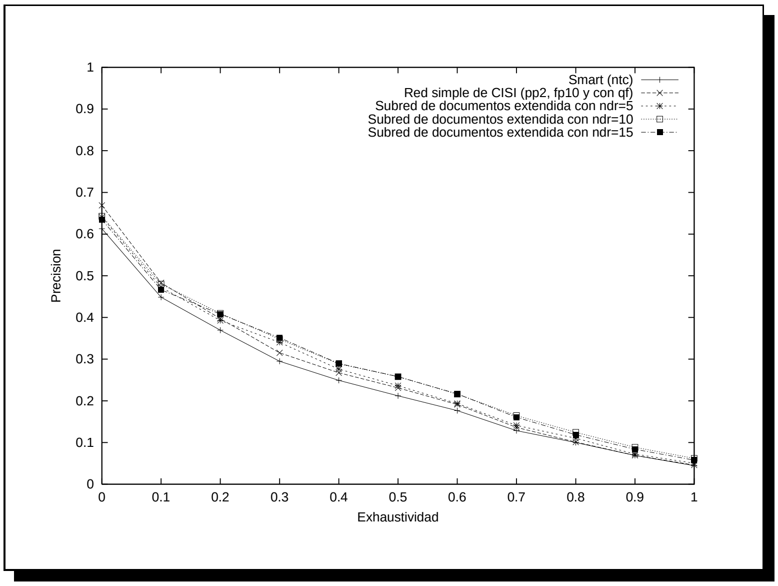


Figura C.3.: Curva Exhaustividad - Precisión para la subred extendida de documentos en la colección CISI.

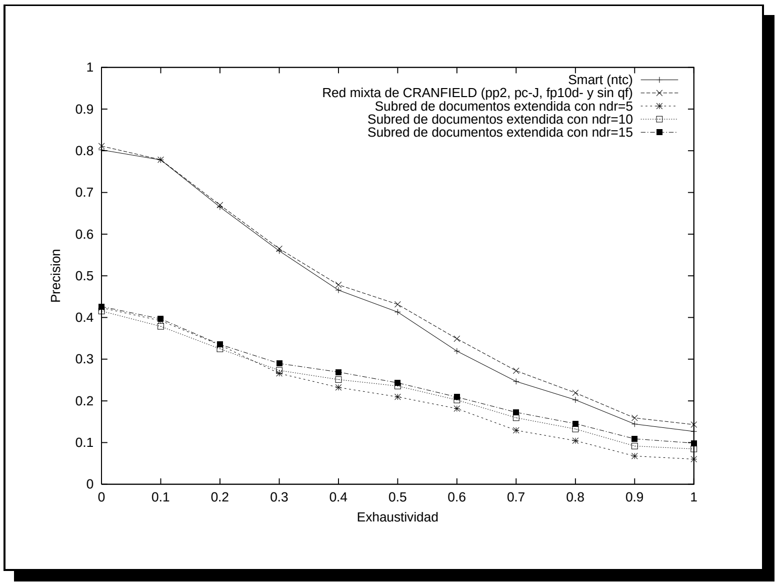


Figura C.4.: Curva Exhaustividad - Precisión para la subred extendida de documentos en la colección CRANFIELD.

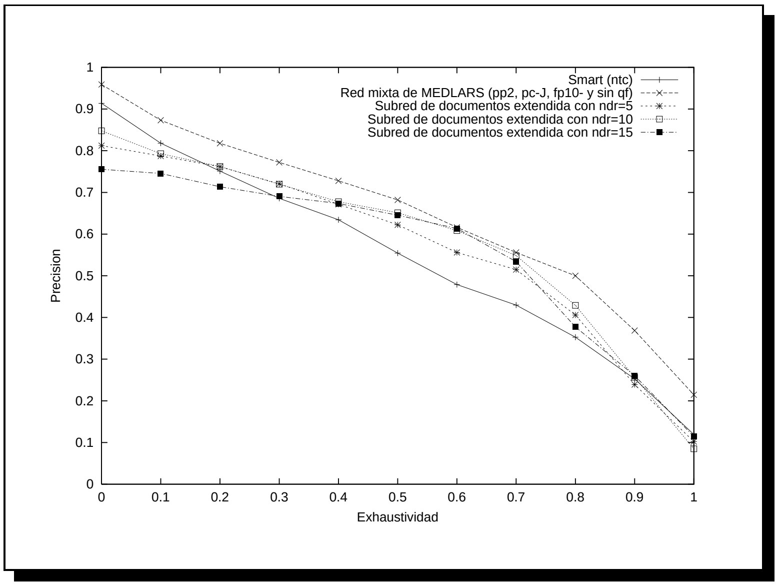


Figura C.5.: Curva Exhaustividad - Precisión para la subred extendida de documentos en la colección MEDLARS.

Bibliografía

[AC01] S. Acid y L. M. de Campos. *A Hybrid methodology for learning belief networks: BENEDICT*. International Journal of Approximate Reasoning (Por aparecer).

[AC96] S. Acid y L. M. de Campos. *Benedict: An algorithm for learning probabilistic belief networks*. En *Proceedings of the Sixth International Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'96)*, páginas 979–984, 1996.

[AC00] S. Acid y L. M. de Campos. *Learning right sized belief networks by means of a hybrid methodology*. Principles of Data Mining and Knowledge Discovery, Lecture Notes in Artificial Intelligence. Springer Verlag, 1910, 309–315, 2000.

[BR99] R. Baeza-Yates y B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, Essex, 1999.

[BC87] J. Belkin y W. B. Croft. *Retrieval techniques*. Annual Review of Information Science and Technology, 22, 109–145, 1987.

[Boo80] A. Bookstein. *Fuzzy request: An approach to weighted boolean searches*. Journal of the American Society for Information Science (JASIS), 31(4), 240–247, 1980.

[Boo83] A. Bookstein. *Outline of a general probabilistic retrieval model*. Journal of Documentation, 399(2), 63–72, 1983.

[Boo79] A. Bookstein. *Relevance*. Journal of the American Society for Information Science (JASIS), 30, 269–273, 1988.

[Bro94] E. W. Brown. *An approach for improving execution performance in inference network based information retrieval*. Technical Report 94–73, University of Massachusetts, 1994.

[BG94b] P. Bruza y L. C. van de Gaag. *Index Expression Belief Networks for information disclosure*. International Journal of Expert Systems, 7(2), 107–138, 1994.

[BI94] P. D. Bruza y J. J. Ijdens. *Efficient probabilistic inference through Index Expression Belief Networks*. En *Proceedings of the Seventh Australian Joint Conference on Artificial Intelligence (AI94)*, páginas 592–599. World Scientific, 1994.

[BG94] M. Buckland y F. Grey. *The relationship between recall and precision*. Journal of the American Society for Information Science (JASIS), 45(1), 12–19, 1994.

[Buc85] C. Buckley. *Implementation of the SMART Information Retrieval System*. Technical Report TR85-686, 1985.

[BSAS94] C. Buckley, G. Salton, J. Allan y A. Singhal. *Automatic query expansion using SMART*. En *Proceedings of the TREC-3 Conference. NIST special publication 500–226*, páginas 69–80, 1994.

[CCB95] J. Callan, W. Croft y J. Broglio. *TREC and TIPSTER experiments with INQUERY*. Information Processing & Management, 31(3), 327–343, 1995.

[CCH92] J. P. Callan, W. B. Croft y S. M. Harding. *The INQUERY Retrieval System*. En *Proceeding of the* 3^{rd} *International Conference on Database and Expert Systems Application. Valencia, Spain, Springer-Verlag*, páginas 78 –83, 1992.

[CLC95] J. P. Callan, Z. Lu y W. B. Croft. *Searching distributed collections with Inference Networks*. En E. A. Fox, P. Ingwersen y F. Raya, editores, *SIGIR'95, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA, July 9-13, 1995 (Special Issue of the SIGIR Forum)*, páginas 21–28. ACM Press, 1995.

[Cam98b] L. M. de Campos. *Aprendizaje automático de modelos gráficos I: métodos básicos*. En Gámez y Puerta [GP98], páginas 113–140.

[Cam98] L. M. de Campos. *Independency relationships and learning algorithms for singly connected networks*. Journal of Experimental and Theoretical Artificial Intelligence, 10(4), 511–549, 1998.

[CFH00] L. M. de Campos, J. M. Fernández y J. F. Huete. *Building Bayesian network-based Information Retrieval systems*. En 11^{th} *International Workshop on Database and Expert Systems Applications:* 2^{nd} *Workshop on Logical and Uncertainty Models for Information Systems (LUMIS)*, páginas 543–552. Database and Expert Systems Applications, 2000.

[CFH98] L. M. de Campos, J. M. Fernández y J. F. Huete. *Query expansion in Information Retrieval systems using a Bayesian network-based thesaurus*. En *Proceedings of the* 14^{th} *Uncertainty in Artificial Intelligence Conference*, páginas 53–60, 1998.

[CH97] L. M. de Campos y J. F. Huete. *On the use of independence relationships for learning simplified belief networks*. Journal of Intelligent Systems, 12, 495–522, 1997.

[CH00] L. M. de Campos y J. F. Huete. *A new approach for learning belief networks using independence criteria*. International Journal of Approximate Reasoning, 24(1), 11–37, 2000.

[CP01] L. M. de Campos y J. M. Puerta. *Stochastic local and distributed search algorithms for learning belief networks*. En *Proceedings of the Third International Symposium on Adaptive Systems (ISAS-2001):Evolutionary Computation and Probabilistic Graphical Models*.

[CP01b] L. M. de Campos y J. M. Puerta. *Stochastic local algorithms for learning Belief networks: Searching in the space of the orderings*. En *Sixth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU-2001) (Por aparecer)*, 2001.

[CGH96] E. Castillo, J. Gutiérrez y A. Hadi. *Sistemas expertos y modelos de redes probabilísticas*. Academia de Ingeniería, 1996.

[Ces90] B. Cestnik. *Estimating probabilities: A crucial task in Machine Learning*. En *Proceedings of European Conference on Artificial Intelligence (ECAI94)*, páginas 147–149, 1990.

[CB91] B. Cestnik y I. Bratko. *On estimating probabilities in tree pruning*. Lecture Notes in Artificial Intelligence, páginas 138–150, 1991.

[CCR71] Y. K. Chang, C. Cirillo y J. Razon. *Evaluation of feedback retrieval using modified freezing, residual collection and test and control groups*. En Salton [Sal71], páginas 355–370.

[CGH94] D. Chickering, D. Geiger y D. Heckerman. *Learning Bayesian networks is NP-hard*. Informe Técnico MSR-TR-94-17, University of California, 1994.

[CL68] C. K. Chow y C. Liu. *Approximating discrete probability distributions with dependence trees*. IEEE Transactions on Information Theory, 14, 462–467, 1968.

[Chr75] N. Christofides. *Graph theory. An algorithmic approach*. Academic Press Inc., London, 1975.

[CH92] G. Cooper y E. Herskovits. *A Bayesian method for the induction of probabilistic networks from data*. Machine Learning, 9, 309–347, 1992.

[Coo90] G. F. Cooper. *Probabilistic inference using belief networks is NP-hard*. Artificial Intelligence, páginas 393–405, 1990.

[CH91] G. F. Cooper y E. Herskovits. *A Bayesian method for constructing Bayesian belief networks from databases*. En *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, páginas 86–94, 1991.

[Coo88] W. S. Cooper. *Getting beyond Boole*. Information Processing & Management, 24, 243–248, 1988.

[CLRC98] F. Crestani, M. Lalmas, C. J. van Rijsbergen y L. Campbell. *Is this document relevant?... probably. A survey of probabilistic models in Information Retrieval*. ACM Computing Survey, 30(4), 528–552, 1991.

[Cro81] W. B. Croft. *Document representation in probabilistic models of information retrieval*. Journal of the American Society for Information Science (JASIS), páginas 451–457, 1981.

[CH79] W. B. Croft y D. J. Harper. *Using probabilistic models of document retrieval without relevance information*. Journal of Documentation, 35(4), 285–295, 1979.

[CT92] W. B. Croft y H. R. Turtle. *Text retrieval and inference*. En L. Erlbaum, editor, *Text-based Intelligent Systems*, páginas 127–156. Paul Jacobs Ed., 1992.

[CT93] W. B. Croft y H. R. Turtle. *Retrieval strategies for hypertext*. Information Processing & Management, 29(3), 313–324, 1993.

[Cro94] V. Cross. *Fuzzy information retrieval*. Journal of Intelligent Information Systems, 3(1), 29–56, 1994.

[DPHS98] S. T. Dumais, J. Platt, D. Hecherman y M. Sahami. *Inductive learning algorithms and representations for text categorization*. En G. Gardarin, J. C. French, N. Pissinou, K. Makki y L. Bouganim, editores, *Proceedings of the 1998 ACM CIKM International Conference on Information and Knowledge Management, Bethesda, Maryland, USA, November 3-7, 1998*, páginas 148–155. ACM, 1998.

[Ell97] D. Ellis. *Progress and problem in information retrieval*. Library Association, London, 1997.

[FF93] B. del Favero y R. M. Fung. *Bayesian inference with aggregation for information retrieval*. En *Proceedings of the TREC-2 Conference*, páginas 151–161, 1993.

[Fox92] C. Fox. *Lexical analysis and stoplist*. En Frakes y Baeza-Yates [FB92], páginas 102–130.

[Fra92] W. B. Frakes. *Introduction to information storage and retrieval*. En Frakes y Baeza-Yates [FB92], páginas 1–12.

[Fra92b] W. B. Frakes. *Stemming algorithms*. En Frakes y Baeza-Yates [FB92], páginas 131–160.

[FB92] W. B. Frakes y R. Baeza-Yates, editores. Prentice-Hall, New Jersey, 1992.

[FC89] M. Frisse y S. B. Cousins. *Information retrieval from hypertext: Update on the Dynamic Medical Handbook Project*. En *Proceedings of the Hypertext'89 Conference. Pittsburgh*, páginas 199–212, 1989.

[Fri88] M. E. Frisse. *Searching for information in a hypertext medical handbook*. Communications of the ACM, 31, 880–886, 1988.

[Fuh89] N. Fuhr. *Models for retrieval with probabilistic indexing*. Information Processing & Management, 25(1), 55–72, 1989.

[Fuh92] N. Fuhr. *Probabilistic models for information retrieval*. The Computer Journal, 35(3), 243–355, 1992.

[FB91] N. Fuhr y C. Buckley. *A probabilistic learning approach for document indexing*. ACM Transactions on Information Systems, 9(3), 223–248, 1991.

[FP94] N. Fuhr y U. Pfeifer. *Probabilistic information retrieval as combination of abstraction inductive learning and probabilistic assumptions*. ACM Transactions on Information Systems, 12(1), 92–115, 1994.

[FF95] R. Fung y B. D. Favero. *Applying Bayesian networks to information retrieval*. Communications of the ACM, 38(2), 42–57, 1995.

[FCAT90] R. M. Fung, S. L. Crawford, L. A. Appelbaum y R. M. Tong. *An architecture for probabilistic concept-based information retrieval*. En J.-L. Vidick, editor, *Proceedings of the SIGIR'90,* 13^{th} *International Conference on Research and Development in Information Retrieval, Brussels, Belgium, 5-7 September 1990*, páginas 455–467. ACM, 1990.

[GPP90] D. Geiger, A. Paz y J. Pearl. *Learning causal trees from dependence information*. En *Proceedings of the* $8^t h$ *National Conference on Artificial Intelligence (AAAI 90)*, páginas 770–776, 1990.

[GPP93] D. Geiger, A. Paz y J. Pearl. *Learning simple causal structures*. International Journal of Intelligent Systems, 8, 231–247, 1993.

[GVP90] D. Geiger, J. Verma y J. Pearl. *Identifying independence in Bayesian networks*. Networks, 20, 507–534, 1990.

[GVP90b] D. Geiger, T. Verma y J. Pearl. *Separation: From theorems to algorithms*. Uncertainty in Artificial Intelligence 5, páginas 139–148, 1990.

[GIS96] D. Ghazfan, M. Indrawan y B. Srinivasan. *Toward meaningful Bayesian networks for information retrieval systems*. En *Proceedings of the IPMU'96 Conference*, páginas 841–846, 1996.

[Glu96] M. Gluck. *Exploring the relationships between user satisfaction and relevance in information systems*. Information Processing & Management, 32(1), 89–104, 1996.

[GP98] J. A. Gámez y J. M. Puerta, editores. Ediciones de la Universidad de Castilla-La Mancha, Cuenca, 1998.

[Goo65] I. J. Good. *The Estimation of Probabilities*. MIT Press, Cambridge, 1965.

[Elv00] D. C. de la Computación e Inteligencia Artificial. Universidad de Granada. *Elvira*. http://leo.ugr.es/ẽlvira/elvira.html, 2000.

[Gre96] W. Greiff. *Computational tractable, conceptually plausible classes of link matrices for the INQUERY Inference Network*. Informe Técnico TR–96–66, University of Massachusetts, 1996.

[GCT97] W. R. Greiff, W. B. Croft y H. R. Turtle. *Computationally tractable probabilistic modeling of Boolean operators*. En *SIGIR '97: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 27-31, 1997, Philadelphia, PA, USA*, páginas 119–128. ACM, 1997.

[HC93] D. Haines y W. B. Croft. *Relevance feedback and inference networks*. En *Proceedings of the* 16^{th} *ACM – SIGIR Conference on research and development in information retrieval*, páginas 2–11, 1993.

[HFC94] C. Han, H. Fujii y W. B. Croft. *Automatic query expansion for Japanese text retrieval*. Informe Técnico IR-57, University of Massachusetts, 1994.

[Har92b] D. Harman. *Relevance feedback and other query modification techniques*. En Frakes y Baeza-Yates [FB92], páginas 241–263.

[Har92] D. Harman. *Relevance Feedback Revisited*. En *Proceedings of the* 16^{th} *ACM–SIGIR Conference on research and development in information retrieval*, páginas 1–10, 1992.

[HFBL92] D. Harman, E. Fox, R. Baeza-Yates y W. Lee. *Inverted files*. En Frakes y Baeza-Yates [FB92], páginas 28–43.

[HR78] D. J. Harper y C. J. van Rijsbergen. *An evaluation of feedback in document retrieval using co-occurrence data*. Journal of Documentation, 34(3), 189–216, 1978.

[Har75a] S. P. Harter. *A probabilistic approach to automatic keyword indexing. Part I. On the distribution of Speciality words in a technical literature*. Journal of the American Society for Information Science (JASIS), 26, 197–205, 1975.

[Har75b] S. P. Harter. *A probabilistic approach to automatic keyword indexing. Part II. An algorithm for probabilistic indexing*. Journal of the American Society for Information Science (JASIS), 26, 280–289, 1975.

[HGC95] D. Heckerman, D. Geiger y D. Chickering. *Learning Bayesian networks: The combination of knowledge and statistical data.* Machine Learning, (20), 197–243, 1995.

[HH98] D. Heckerman y E. Horvitz. *Inferring informational goals from free-text queries: A Bayesian approach.* En *Proceedings of the* 14^{th} *Uncertainty in Artificial Intelligence Conference*, páginas 230–237, 1998.

[Her98] L. D. Hernández. *Algoritmos de progación I. Métodos exactos de inferencia*. En Gámez y Puerta [GP98], páginas 41–64.

[HC90] E. Herskovits y G. Cooper. *Kutató: An entropy-driven system for the construction of probabilistic expert systems from databases*. En *Conference on Uncertainty in Artificial Intelligence*, páginas 54–62, 1990.

[HC93b] J. F. Huete y L. M. de Campos. *Learning causal polytrees*. En *Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, páginas 180–185. Lecture Notes in Computer Science 747. Eds M. Clarke and R. Kruse and S. Moral, 1993.

[Ijd94] J. J. Ijdens. *Using Index Expression Belief Network for information disclosure. Toward effective use of the IEBN model as a disclosure system. Master thesis. Department of Computer Science. Utrecht University*, 1994.

[IBH95] J. J. Ijdens, P. D. Bruza y D. J. Harper. *Probabilistic inference experiments using ECLAIR Framework.* Technical Report 95/7, 1995.

[IGS96] M. Indrawan, D. Ghazfan y B. Srinivasan. *Using Bayesian networks as retrieval engines*. En *Proceedings of the* 6^{th} *Text Retrieval Conference. NIST Special Publication 500-238*, páginas 437–444. NIST, 1996.

[JMF99] A. K. Jain, M. Murty y P. J. Flynn. *Data clustering: a review*. ACM Computing Surveys, 31(3), 264–323, 1999.

[JLO90] F. Jensen, S. Lauritzen y K. Olesen. *Bayesian updating in causal probabilistic networks by local computations*. Computational Statistics Quarterly, 5(4), 269–282, 1990.

[Jen96] F. V. Jensen. *An Introduction to Bayesian Networks*. University College London Press, London, 1996.

[JC94] Y. Jing y W. B. Croft. *An association thesaurus for information retrieval*. En *Proceedings of the RIAO'94 conference. New York*, páginas 146–160, 1994.

[Jon79] K. S. Jones. *Search term relevance weighting given very little relevance information*. Journal of Documentation, 35(1), 30–48, 1979.

[Kor97] R. R. Korfhage. *Information storage and retrieval*. John Wiley and son, Inc., 1997.

[KBP99] D. Kraft, P. Bordogna y G. Pasi. *Fuzzy set techniques in information retrieval*. En Fuzzy Set Techniques in Information Retrieval. Didier, D. and Prade, H. (eds.) Handbook of Fuzzy Sets and Possibility Theory. Approximate Reasoning and Fuzzy Information Systems. Kluwer Academic Publishers, AA Dordrecht, The Netherlands, Chp. 8., 1999.

[Kul68] S. Kullback. *Information theory and statistics*. Dover Publications, 1968.

[KL51] S. Kullback y R. Leibler. *On information and sufficiency*. Annals of Mathematical Statistics, (22), 76–86, 1951.

[LS88] S. Lauritzen y D. Spiegelhalter. *Local computations with probabilities on graphical structures and their applications to expert systems (with discussion)*. The Journal of the Royal Statistical Society, 50, 157–224, 1988.

[MK60] M. E. Maron y J. L. Kuhns. *On relevance, probabilistic indexing and information retrieval*. Journal of the Association for Computer Machinery, 7, 216–244, 1960.

[Miz97] S. Mizzaro. *Relevance: The whole history*. Journal of the American Society for Information Science (JASIS), 48(9), 810–832, 1997.

[MCKH90] H. Mori, C. Chung, Y. Kinoe y Y. Hayashi. *An adaptative document retrieval system using a neural network*. International Journal of Human-Computer Interaction, 2(3), 267–280, 1990.

[Nea90] R. E. Neapolitan. *Probabilistic reasoning in Expert Systems. Theory and algorithms*. John Wiley and Sons, 1990.

[ORM91] M. P. Oakes, D. Reid y t. McEnery. *Some practical applications of neural networks in information retrieval*. En 13^{th} *Information Retrieval Colloquium, 8–9 April, Lancaster (UK)*, páginas 167–187. British Computer Society, 1991.

[Pao78] M. L. Pao. *Automatic text analysis based on transition phenomena of word occurrences*. Journal of American Society for Information Science (JASIS), páginas 121–124, 1978.

[Par97] H. Park. *Relevance of science information: Origins and dimensions of relevance and their implications to information retrieval*. Information Processing & Management., 33(3), 339–352, 1997.

[PC96] Y. Park y K. Choi. *Automatic thesaurus construction using Bayesian networks*. Information Processing & Management, 32(5), 543–553, 1996.

[Pea86] J. Pearl. *A constraint-propagation approach to probabilistic reasoning*. En L. Kanal y J. Lemmer, editores, *Proceedings of the Uncertainty in Artificial Intelligence conference*, páginas 357–370. North-Holland, Amsterdam, 1986.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan and Kaufmann, San Mateo, California, 1988.

[Pea93] J. Pearl. *Belief networks revisited*. Informe Técnico R-175, Computer Science Department, University of California, 1993.

[Por80] M. F. Porter. *An algorithm for suffix stripping*. Program, 14, 130–134, 1981.

[Ras92] E. Rasmussen. *Clustering algorithms*. En Frakes y Baeza-Yates [FB92], páginas 419–442.

[RP89] G. Rebane y J. Pearl. *The recovery of causal polytrees from statistical data*. En L. Kanal, T. S. Levitt y J. F. Lemmer, editores, *Uncertainty in Artificial Intelligence*, páginas 175–182. North Holland, 1989.

[RM96] B. A. Ribeiro-Neto y R. R. Muntz. *A Belief network model for IR*. En H. Frei, D. Harman, P. Schäble y R. Wilkinson, editores, *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)*, páginas 253–260. ACM, 1996.

[Rij77] C. J. van Rijsbergen. *A theoretical basis for the use of co-occurrence data in information retrieval*. Journal of Documentation, 33(22), 106–119, 1977.

[Rij79] C. J. van Rijsbergen. *Information retrieval. Second Edition*. Butter Worths, London (U.K.), 1979.

[RHP81] C. J. van Rijsbergen, D. J. Harper y M. F. Porter. *The selection of good search terms.* Information Processing & Management, 17, 77–91, 1981.

[Rob77] S. E. Robertson. *The probability ranking principle in IR*. Journal of Documentation, 33(4), 294–304, 1977.

[RJ76] S. E. Robertson y K. S. Jones. *Relevance weighting of search terms*. Journal of the American Society for Information Science, 27(3), 129–146, 1976.

[RRP80] S. E. Robertson, C. J. van Rijsbergen y M. E. Porter. *Probabilistic models of indexing and searching*. En *Information Retrieval Research, Proc. Joint ACM/BCS Symposium in Information Storage and Retrieval, Cambridge, June 1980*, páginas 35–56. Butterworths, London, 1980.

[Sah96] M. Sahami. *Learning Limited Dependence Bayesian Classifiers*. En *Second International Conference on Knowledge Discovery and Data Mining in Databases*, páginas 335–338, 1996.

[Sah98] M. Sahami. *Using machine learning to improve information access.* Tesis Doctoral, Stanford University, 1998.

[SDHH98] M. Sahami, S. Dumais, D. Heckerman y E. Horvitz. *A Bayesian approach to filtering junk e-mail*. En *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[SYB98] M. Sahami, S. Yusufali y M. Q. W. Baldonado. *SONIA: A Service for Organizing Networked Information Autonomously*. En *Proceedings of the* 3*rd* *ACM International Conference on Digital Libraries, June 23-26, 1998, Pittsburgh, PA, USA*, páginas 200–209. ACM Press, 1998.

[Sal98] A. Salmerón. *Algoritmos de progación II. Métodos de Monte Carlo*. En Gámez y Puerta [GP98], páginas 65–88.

[SCM00] A. Salmerón, A. Cano y S. Moral. *Importance sampling in Bayesian networks using probability trees*. Computational Statistics & Data analysis, (34), 387–413, 2000.

[Sal71] G. Salton, editor. Prentice Hall, Inc., Englewood Cliffs, NJ, 1971.

[SB88] G. Salton y C. Buckley. *Term-weighting approaches in automatic text retrieval*. Information Processing & Management, 24, 513–523, 1988.

[SB90] G. Salton y C. Buckley. *Improving retrieval performance by relevance feedback*. Journal of the American Society for Information Science (JASIS), 41, 288–297, 1990.

[SL68] G. Salton y M. E. Lesk. *Computer evaluation of indexing and text precising*. Journal of the Association of Computing Machinery, 15, 8–36, 1968.

[SM83] G. Salton y M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1983.

[SWY75] G. Salton, A. Wong y C. S. Yang. *A vector space model for automatic indexing*. Communications of the ACM, 18(11), 613–620, 1975.

[Sar75] T. Saracevic. *Relevance: A review of and a framework for thinking on the notion in information science.* Journal of the American Society for Information Science (JASIS), 26, 321–343, 1975.

[SKCT88] T. Saracevic, P. Kantor, A. Y. Chamis y D. Trivison. *A study of information seeking and retrieving. I. Background and methodology*. Journal of the American Society for Information Science (JASIS), 39, 161–176, 1988.

[Sav92] J. Savoy. *Bayesian inference networks and spreading activation in hypertext systems*. Information Processing & Management, 28(3), 389–406, 1992.

[Sav93] J. Savoy. *Searching for information in hypertext systems using multiple sources of evidence*. International Journal of Man-Machine Studies, 38, 1017–1030, 1993.

[Sav95] J. Savoy. *An evaluation of probabilistic retrieval models*. Technical Report CR-I-95-05, 1995.

[SD91] J. Savoy y D. Desbois. *Information retrieval in hypertext systems: an approach using Bayesian networks*. Electronic publishing, 4(2), 87–108, 1991.

[SAS94] R. Schacter, S. Andersen y P. Szolovits. *Global conditioning for probabilistic inference in belief networks*. En *Proceedings of the* 10^{th} *Uncertainty in Artificial Intelligence Conference*, páginas 514–522. Morgan Kaufmann Publisher, 1994.

[Sch93] J. Scholtes. *Neural Networks in Natural Language Processing and Information Retrieval*. Tesis Doctoral, Universiteit van Amsterdam, Amsterdam,the Netherlands., 1993.

[SP97] H. Schütze y J. Pedersen. *A coocurrence-based thesaurus and two applications to information retrieval*. Information Processing & Management, 33(3), 307–318, 1997.

[SRCMZ00] I. Silva, B. Ribeiro-Neto, P. Calado, E. Moura y N. Ziviani. *Link-based and content-based evidential information in a belief network model*. En *Proceedings of the* 23^{th} *International ACM–SIGIR Conference on Research and Development in Information Retrieval, Athens, July 2000*, páginas 96–103. ACM, 2000.

[Rei00] I. R. Silva. *Bayesian Networks for Information Retrieval Systems.* Tesis Doctoral, Universidad Federal de Minas Gerais, 2000.

[SL96] A. Spink y R. M. Losee. *Feedback in Information Retrieval*. Annual Review of Information Science and Technology (ARIST), 31, 33–78, 1996.

[SGS91] P. Spirtes, C. Glymour y R. Scheines. *An algorithm for fast recovery of sparse causal graphs*. Social Science Computer Review, 9, 62–72, 1991.

[SGS93] P. Spirtes, C. Glymour y R. Scheines. *Causation, Prediction and Search*. Lecture Notes in Statistics 81. Springer Verlag, New York, 1993.

[Sri92] P. Srinivasan. *Thesaurus construction*. En Frakes y Baeza-Yates [FB92], páginas 161–218.

[Su94] L. T. Su. *The relevance of recall and precision in user evaluation*. Journal of the American Society for Information Science (JASIS), 45(3), 207–217, 1994.

[SC91b] H. Suermondt y G. Cooper. *A combination of exact algorithms for inference in Bayesian belief networks*. International Journal of Approximate Reasoning, 5, 83–94, 1991.

[SC91] H. Suermondt y G. Cooper. *Initialization for the Method of Conditioning in Bayesian Belief Networks*. Artificial Intelligence, 40(1), 83–94, 1991.

[SCH91] H. Suermondt, G.Cooper y D.Heckerman. *A combination of cutset conditioning with clique-tree propagation in the Pathfinder system*. En *Proceedings of the* 6^{th} *Conference on Uncertainty in Artificial Intelligence*, páginas 273–279, 1990.

[TC92] H. Turtle y W. B. Croft. *A Comparison of text retrieval models*. The Computer Journal, 35(3), 279–290, 1992.

[Tur90] H. R. Turtle. *Inference networks for document retrieval*. Tesis Doctoral, Universidad de Massachusetts, 1990.

[TC91] H. R. Turtle y W. B. Croft. *Efficient probabilistic inference for text retrieval*. En *Proceedings of the RIA0'91 Conference. Barcelona (España)*, páginas 644–661, 1991.

[TC91b] H. R. Turtle y W. B. Croft. *Evaluation of an Inference network-Based retrieval model*. Information Systems, 9(3), 187–222, 1991.

[TC97] H. R. Turtle y W. B. Croft. *Uncertainty in information retrieval systems*. En A. Motro y P. Smets, editores, *Uncertainty management in information systems: from needs to solutions*, páginas 189–224. Kluwer Academic Publishers, 1997.

[TH93] K. Tzeras y S.Hartmann. *Automatic Indexing Based on Bayesian Inference Networks*. En R. Korfhage, E. M. Rasmussen y P. Willett, editores, *Proceedings of the* 16^{th} *Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Pittsburgh, PA, USA, June 27 - July 1, 1993*, páginas 22–34. ACM, 1993.

[VP90] T. Verma y J. Pearl. *Causal Networks: Semantics and expressiveness*. En R. D. Shachter, T. S. Lewitt, L. Kanal y J. F. Lemmer, editores, *Proceedings of the Uncertainty in Artificial Intelligence Conference*, páginas 69–76. North-Holland, 1990.

[War92] S. Wartik. *Boolean operations*. En Frakes y Baeza-Yates [FB92], páginas 264–292.

[WMB99] I. H. Witten, A. Moffat y T. C. Bell. *Managing gigabytes. Compressing and indexing documents and images.* 2^{nd} *edition*. Morgan Kaufman Publishers, Inc., San Francisco, 1999.

[XC96] J. Xu y W. B. Croft. *Query Expansion Using Local and Global Document Analysis*. En H. Frei, D. Harman, P. Schäble y R. Wilkinson, editores, *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)*, páginas 4–11. ACM, 1996.

[XC00] J. Xu y W. B. Croft. *Improving the Effectiveness of information retrieval systems with local context analysis*. ACM Transactions on Information Systems, 18(1), 79–112, 2000.

[Zad65] L. Zadeh. *Fuzzy Sets*. Information and Control, 8(3), 338–353, 1965.