Curso de Doctorado

"Representación del Conocimiento"

Tratamiento de la incertidumbre en sistemas de recuperación de información: Recuperación Inteligente de Información

Fernando Berzal Galiano

Diseño, Análisis y Aplicaciones de Sistemas Inteligentes

Departamento de Ciencias de la Computación e Inteligencia Artificial

Universidad de Granada

Índice

Introducción

Tradicionalmente, los ordenadores se han aplicado a la resolución de problemas en los cuales los datos estaban perfectamente estructurados (vg: problemas de cálculo numérico). No obstante, hoy en día existe una gran cantidad de información textual con la que hay que tratar (revistas, periódicos, libros, informes, ensayos...). El procesamiento de textos en formato libre es esencial en la "era de la información".

El objetivo de un sistema de recuperación de información [IRS: Information Retrieval System] es buscar en un almacén de información, tal como una base de datos documental o la World Wide Web, y obtener documentos potencialmente relevantes ante una consulta dada. Ya que estos sistemas han de operar en tiempo real, deben buscar en grandes volúmenes de datos con rapidez y eficiencia.

Usualmente, se permite al usuario que busque documentos que contengan palabras clave o expresiones, las cuales pueden combinarse empleando operadores booleanos. También se han utilizado estimaciones probabilísticas para determinar si un documento satisface la consulta del usuario (vg: redes bayesianas). Para mejorar su rendimiento, algunos sistemas permiten una realimentación por parte del usuario, el cual ha de indicar cuáles de los documentos recuperados son realmente relevantes. Esa información se utiliza para modificar la consulta original, añadiendo nuevos términos [automatic query expansion] o modificando sus pesos en la consulta original.

Los sistemas de recuperación de información deben tratar adecuadamente las ambigüedades e idiosincrasias del lenguaje natural, como la sinonimia (palabras con el mismo significado) o la polisemia (un mismo término puede tener distintas acepciones). Además, las expresiones requieren una atención especial porque su significado no siempre corresponde al de las palabras individuales que las constituyen (vg: "sistema operativo", en Informática, no es simplemente un sistema que esté funcionando). Por tanto, se hace recomendable la utilización de técnicas de procesamiento del lenguaje natural [NLP: natural language processing] en la mejora de los sistemas actuales de recuperación de información.

Algunos sistemas no se limitan únicamente a indexar documentos para responder eficientemente a las consultas de los usuarios, sino que también intentan extraer información para generar resúmenes, analizar relaciones entre términos, etc.

Otras aplicaciones interesantes relacionadas con los sistemas de recuperación de información incluyen la clasificación automática de documentos o su filtrado. La clasificación automática de documentos [text categorization] puede emplearse para automatizar la construcción de grandes índices de información en Internet. La segunda de las aplicaciones mencionadas podría utilizarse, por ejemplo, para reducir las molestias causadas por el correo electrónico indeseado [spam]. Una tercera aplicación, la selección de documentos que puedan resultar interesantes a un usuario dado su perfil [text routing] tampoco carece de interés práctico.

Propedeútica

La propedéutica se define como la "enseñanza preparatoria para el estudio de una disciplina" en el Diccionario de la Real Academia Española de la Lengua. Este apartado no es más que eso, una recopilación de conceptos necesarios para el estudio de los Sistemas de Recuperación de Información.

La mayor parte de los sistemas de recuperación de información preprocesan los documentos para construir un índice invertido que les permite determinar eficientemente en qué documentos aparece una palabra determinada. Además, las palabras más comunes en un lenguaje (como determinantes y preposiciones) no se suelen considerar al construir los índices ya que, en principio, no contribuyen en gran medida al significado de los documentos [stopwords]. Por otra parte, se suelen emplear algoritmos que intentan extraer el lexema de una palabra con el objetivo de agrupar las distintas variaciones morfológicas de un mismo término [stemming], tales como formas verbales, singular/plural o masculino/femenino.

Una representación alternativa de los documentos consiste en emplear firmas [signatures] o n-gramas. De esta forma, el documento se reduce a una codificación binaria de longitud fija. Dicha codificación puede resultar particularmente atractiva para su implementación en sistemas paralelos o hardware de propósito especial (como las arquitecturas sistólicas), pero puede producir resultados totalmente erróneos al suponer la aparición en un documento de términos que no están en él realmente.

Para evaluar la eficacia de un sistema de recuperación de información hacen falta medidas estandarizadas. Obviamente, dos factores esenciales son el tiempo de respuesta y la facilidad de uso. El primer factor se puede medir directamente, aunque el segundo es bastante subjetivo. La efectividad de un sistema de recuperación de información se suele estimar utilizando dos medidas de relevancia para las consultas en función de los documentos que obtienen: la retentiva [recall] y la precisión [precision]. La retentiva es la proporción de documentos relevantes recuperados respecto al número total de documentos relevantes mientras que la precisión es la proporción de documentos relevantes recuperados respecto al total de los documentos recuperados (relevantes e irrelevantes).

Un sistema efectivo de recuperación de información debería incluir técnicas como:

- *Operadores booleanos*: El operador AND se puede emplear para mejorar la precisión (incrementar el número de documentos relevantes minimizando el número de documentos irrelevantes recuperados) mientras que el operador OR permite mejorar la retentiva del sistema (ampliando el número de documentos relevantes recuperados).

- *Uso de pesos*: El uso de pesos para los términos que aparecen en un documento puede mejorar la precisión de un sistema de recuperación de información.

- *Análisis morfológico*: La eliminación de prefijos y sufijos, así como la reducción de las palabras a sus lexemas, puede contribuir a la mejora de la retentiva del sistema.

- *Métodos estadísticos*: La eliminación de palabras comunes [stop words], la contabilización de la frecuencia de aparición de los distintos términos y el análisis de los términos que aparecen en los documentos relativos a un tema permiten la construcción de perfiles para los documentos.

- *Métodos heurísticos*: Técnicas como el análisis de las referencias bibliográficas [bibliographic coupling & co-citation analysis] se pueden emplear para crear perfiles para los documentos y también perfiles para los usuarios. Se pueden idear reglas que permitan la identificación de referencias a personas, lugares o fechas, las cuales pueden aparecer en distintos formatos aun haciendo referencia a una misma entidad.

- *Tesauros*: El uso de sinónimos en el lenguaje natural es algo que debería tener en cuenta un buen sistema de recuperación de información.

- *Refinamiento de las consultas*: Utilizando la información suministrada por el usuario (que indica cuáles de los documentos recuperados son realmente relevantes) se pueden elaborar perfiles que reduzcan el número de documentos irrelevantes recuperados en futuras consultas.

- *Ordenación de los resultados* [hit result ranking]: Cuando se obtiene una gran cantidad de documentos potencialmente relevantes es aconsejable ordenarlos de forma que primero aparezcan los que puedan tener un mayor interés para el usuario. Para ello se podría utilizar, por ejemplo, la frecuencia de aparición de los términos incluidos en la consulta.

Modelos de Recuperación de Información

Un sistema general de recuperación de información se puede modelar de la siguiente forma (Kraft, 1985):

D: Conjunto de documentos

T: Conjunto de términos que aparecen en el índice

F: Función de indexación, donde F: DxT → [0,1]

Q: Conjunto de consultas del usuario

a: Función de peso para las consultas, donde a: QxT → [0,1]

g: Función de emparejamiento para un único término, g: [0,1]x[0,1]→[0,1]

e: Función de emparejamiento para múltiples términos, e: [0,1]* → [0,1]

La función de indexación F relaciona los documentos con los términos en el índice. F(d,t)=0 implica que el documento d no tiene nada que ver con el/los concepto(s) representado(s) por el término t, mientras que F(d,t)=1 indica que el documento está perfectamente representado por dicho(s) concepto(s).

La función g(F(d,t),a(q,t)) nos da la estimación de la relevancia de un documento 'd' dada una consulta 'q' de un único término 't'. Si la consulta incluye más de un término, la función g puede interpretarse como la evaluación del documento en cuestión respecto a uno de los términos de la consulta.

Finalmente, la función 'e' proporciona la estimación de la relevancia de un documento 'd' respecto a una consulta de n términos, donde cada término es evaluado independientemente con la función 'g'.

Los distintos modelos de recuperación de información pueden derivarse de este modelo general imponiendo restricciones e interpretando sus distintos componentes:

Modelos booleanos

Si se restringe la función F a los valores {0,1} se obtiene el caso típico de indexación clásica, en el cual un término está en un documento (1) o no (0). Del mismo modo, en el modelo booleano, la función de peso 'a' está restringida al conjunto {0,1}. La función 'e', por su parte, mantiene la semántica de los operadores booleanos (AND, OR y NOT).

Permitiendo que la función F tome valores en el intervalo [0,1] se puede establecer un orden en los documentos recuperados como respuesta a una consulta, de forma que se mejora el rendimiento del sistema.

Modelos basados en conjuntos difusos

Se puede ver la función F como una función de pertenencia en un conjunto difuso: el grado en que un documento d pertenece al conjunto de documentos relativos al concepto representado por el término t. Así mismo, se puede evaluar la correspondencia entre una consulta y un documento usando operadores difusos (t-normas, t-conormas y operadores promedio).

Modelos probabilísticos

Otra interpretación de la función 'e' es considerarla la probabilidad condicionada de que un documento sea relevante dada una consulta relativa a los conceptos representados por sus términos.

El modelo probabilístico de recuperación de información más general parte del principio PRP [Probabilistic Ranking Principle] de Robertson (1977): "Para un comportamiento óptimo, un sistema [de recuperación de información] debería ordenar los documentos de acuerdo a la probabilidad de que sean juzgados relevantes o útiles para el problema o necesidad del usuario".

Una cuestión que se deriva del anterior principio es precisamente cuál es el significado de esa probabilidad de relevancia y qué información ha de usarse para estimarla. Distintas respuestas a esta cuestión han dado lugar a distintas variaciones del modelo probabilístico:

- MODELO 1: Se estima la probabilidad de relevancia de un documento individual respecto a una clase de consultas.

- MODELO 2: Se estima la probabilidad de relevancia de una clase de documentos respecto a una consulta concreta.

- MODELO 3: Intenta unificar los dos modelos anteriores. Pretende capturar información histórica acerca de la relevancia de los documentos individuales en relación con una clase de consultas previas y combinar esta información con la información actual acerca de la relevancia de una clase de documentos ante una consulta dada (también denominada necesidad de información).

VSM [Vector Space Model]

Tanto los documentos como las consultas se representan como vectores en un espacio vectorial real, [0,1]^t, de conceptos representados por los términos del índice. Dada una consulta, los documentos se ordenan en función de la similaridad entre sus representaciones vectoriales y el vector correspondiente a la consulta.

La función de similaridad más utilizada es el coseno del ángulo entre los vectores de la consulta y del documento:

$$\cos(q,d) = \frac{q \cdot d}{|q||d|}$$

donde q·d es el producto escalar de los vectores y || es la norma asociada a dicho producto (es decir, |v|=sqrt(v·v)). Por tanto, la fórmula anterior puede verse como una versión normalizada del producto escalar.

La versión original del modelo vectorial asumía que la función 'a' valía 0 ó 1 para cada término de la consulta. Posteriormente, se introdujo la posibilidad de introducir pesos en el intervalo [0,1].

La importancia de un término en un documento

Hay distintas formas de estimar la función 'F' del modelo general de sistema de recuperación de información. Salton (1989) sugirió emplear una medida conocida como **IDF** [inverse document frequency]:

$$d.idf_t = d.tf_t \cdot \log \frac{N}{N_t}$$

donde $d.tf_t$ es el número de veces que el término t aparece en el documento d (**TF** [Term Frequency]), N es el número de documentos en la colección y N_t es el número de documentos de la colección en los que el término t aparece al menos una vez.

Para utilizar la medida anterior con operadores booleanos extendidos, cuyos argumentos toman valores en el intervalo [0,1], ésta se ha de normalizar:

$$d.w_t = \frac{d.tf_t}{\max d.tf_t} \cdot \frac{d.idf_t}{\max d.idf_t}$$

E.A. Fox propuso un esquema alternativo en su tesis doctoral en 1983:

$$d.w_t = \begin{cases} (0.5 + 0.5\frac{d.tf_t}{\max d.tf_t})\frac{\log(N/n_t)}{\log N} & si\ d.tf_t > 0 \\ 0 & si\ d.tf_t = 0 \end{cases}$$

El modelo INQUERY, por su parte, utiliza:

$$d.w_t = \begin{cases} (0.4 + 0.6\cdot(0.4 + 0.6\frac{\log(d.tf_t + 0.5)}{\log(\max d.tf_t + 1.0)}))\cdot\frac{\log(n/n_t)}{\log N} & si\ d.tf_t > 0 \\ 0.4 & si\ d.tf_t = 0 \end{cases}$$

donde los pesos, sin embargo, quedan en el intervalo [0.4,1).

En las tres variantes expuestas el peso de un término en un documento es mínimo cuando su frecuencia de aparición es 0 y máximo cuando es el documento en que más veces aparece.

La incertidumbre en los sistemas de recuperación de información

La indexación de los documentos en un sistema de recuperación de información sólo nos ofrece un conocimiento parcial acerca del contenido de los documentos. No se puede esperar que un sistema clásico identifique inequívocamente los documentos de interés para el usuario. Cualquier medida de relevancia que se establezca tomando como punto de partida la información contenida en los índices será, por tanto, imprecisa.

La estrategia de búsqueda que emplea operadores booleanos estándar (AND, OR y NOT) se suele considerar demasiado restrictiva. Aún así, es la utilizada en la mayor parte de los sistemas comerciales. También se pueden utilizan otros modelos para ordenar los documentos en función del grado de similaridad entre los documentos y la consulta formulada por el usuario (vg: los operadores booleanos extendidos están basados en la Teoría de los Conjuntos Difusos).

En general, se presupone que la consulta formulada por el usuario refleja con precisión lo que el usuario solicita. En la práctica esto no es siempre así, por lo que habrá que incluir técnicas que faciliten el refinamiento de las consultas y la realimentación por parte del usuario. Es más, el concepto de relevancia para el usuario es algo impreciso de por sí.

Operadores booleanos extendidos

La salida de los sistemas de recuperación de información es un conjunto de referencias a documentos. Dichas referencias le muestran al usuario documentos potencialmente interesantes. Sin embargo, cuando el conjunto de referencias obtenido es extenso, el sistema de recuperación de información debería indicar cuáles de ellas es más probable que sean de interés para el usuario. De esta forma, al mostrarle al usuario una secuencia de documentos ordenados de forma decreciente de acuerdo a una función de similaridad entre su consulta y cada documento, se pretende minimizar el tiempo que usuario ha de emplear para encontrar información útil.

Los sistemas booleanos de recuperación de información han sido los más empleados por su eficiencia y la facilidad con la que se realizan consultas. Sin embargo, el modelo booleano clásico no permite ordenar los documentos de mayor a menor similaridad con la consulta realizada por el usuario. Los modelos basados en conjuntos difusos y el modelo booleano extendido son extensiones del modelo booleano básico que superan esta limitación.

Un sistema de recuperación de información basado en operadores booleanos extendidos (sin considerar pesos en los términos de las consultas) viene dado por $<T,Q,D,F>$, donde

T es el conjunto de términos que aparecen en el índice y se utilizan para representar los documentos y las consultas.

Q es el conjunto de consultas admitidas por el sistema. Cada consulta $q \in Q$ es una expresión booleana compuesta por términos de índice y los operadores lógicos AND, OR y NOT.

D es el conjunto de documentos. Cada documento $d \in D$ se representa por $\{(t_1, d.w_1), ..., (t_n, d.w_n)\}$ donde $d.w_i \in [0,1]$ representa el peso del término t_i en el documento d.

F es la función de recuperación de información $F: DxT \rightarrow [0,1]$. La función F asigna a cada par (d,q) un valor entre cero y uno que es una medida de similaridad entre el documento d y la consulta q. El valor de F es el valor del documento d respecto a la consulta q y permite ordenar los documentos de acuerdo a su posible importancia. La función de recuperación F se evalúa de la siguiente forma:

1. Para cada término ti de una consulta, la función $F(d,t_i)$ es el peso del término t_i en el documento d: $d.w_i$.

2. Los operadores lógicos se evalúan aplicando las correspondientes fórmulas, algunas de las cuales se comentarán a continuación.

   NOTA: El operador NOT se evalúa en todos los modelos tratados en este trabajo como: $F(d, NOT\ t_i) = 1 - d.w_i$.

Ejemplos:

Modelo basado en conjuntos difusos (F)

$$F(d, t_1\ AND\ t_2) = \min\{F(d,t_1), F(d,t_2)\}$$
$$F(d, t_1\ OR\ t_2) = \max\{F(d,t_1), F(d,t_2)\}$$
$$F(d,\ NOT\ t_1) = 1 - F(d,t_1)$$

Modelo booleano extendido (E)**:** ***p-normas***

$$F(d, t_1\ AND\ t_2) = 1 - \sqrt[p]{\frac{(1 - F(d,t_1))^p + (1 - F(d,t_2))^p}{2}}$$
$$F(d, t_1\ OR\ t_2) = \sqrt[p]{\frac{F(d,t_1)^p + F(d,t_2)^p}{2}}$$
$$F(d,\ NOT\ t_1) = 1 - F(d,t_1)$$

Los dos modelos expuestos arriba presentan algunos problemas. El modelo basado en conjuntos difusos ha sido criticado porque en determinadas ocasiones genera los resultados ordenados de una forma poco adecuada (los operadores min y max tienen propiedades que afectan negativamente a la efectividad del sistema de recuperación de información). Aunque el modelo booleano extendido no presenta ese problema, el coste computacional de los operadores es elevado.

La Teoría de los Conjuntos Difusos de Zadeh (1965) es una generalización de la Teoría de Conjuntos clásica en la cual se pueden aplicar los conceptos definidos sobre ésta a conjuntos de objetos cuya función de pertenencia varía en el intervalo [0,1].

Se han propuesto distintos operadores difusos como sustitutos de los operadores clásicos AND y OR. Esos operadores se clasifican normalmente en operadores T (T-normas [= normas triangulares] y T-conormas, también llamadas S-normas) y en operadores promedio A (vg: operador OWA). El mínimo es una T-norma mientras que el máximo es una T-conorma.

Los operadores T se han utilizado para modelar decisiones utilizando la intersección o la unión de conjuntos difusos. Sin embargo, hay decisiones humanas que no se adaptan bien a los operadores T; de ahí la aparición de los operadores promedio, cuyo resultado se controla mediante parámetros.

Operadores T

	T-norma [AND]	*T-conorma [OR]*	
F	$\min\{x, y\}$	$\max\{x, y\}$	*"fuzzy set"*
T_1	$x \cdot y$	$x + y - xy$	*INQUERY boolean*
T_2	$\max\{x + y - 1, 0\}$	$\min\{x + y, 1\}$	
T_3	$\frac{xy}{x + y - xy}$	$\frac{x + y - 2xy}{1 - xy}$	
T_4	$\begin{cases} x & si\ y = 1 \\ y & si\ x = 1 \\ 0 & en\ otro\ caso \end{cases}$	$\begin{cases} x & si\ y = 0 \\ y & si\ x = 0 \\ 1 & en\ otro\ caso \end{cases}$	

NOTA: Aunque aquí no se recojan, también se han propuesto operadores T parametrizados.

Operadores A

A_1	$(x + y - xy)^{\gamma}(xy)^{1-\gamma}$	
A_2	$\gamma \max\{x, y\} + (1 - \gamma)\min\{x, y\}$	*Waller-Kraft*
A_3	$\gamma(x + y - xy) + (1 - \gamma)(xy)$	*"network Boolean"*
$A_{4,AND}$	$\gamma \min\{x, y\} + \frac{(1 - \gamma)(x + y)}{2}$	*"infinite-one"*
$A_{4,OR}$	$\gamma \max\{x, y\} + \frac{(1 - \gamma)(x + y)}{2}$	

NOTA: El comportamiento de los operadores promedio mostrados se controla mediante el parámetro γ entre 0 y 1 (vg: 0.3).

Los operadores min y max del modelo F generan una ordenación de los documentos no acorde con la intuición humana porque su resultado depende únicamente de uno de sus operandos (no importa cuál):

d_1 = { (recuperación,0.40), (información,0.40) }
d_2 = { (recuperación,0.99), (información,0.39) }

q = recuperación AND información

F (d_1,q) = min{0.40,0.40} = 0.40
F (d_2,q) = min{0.99,0.39} = 0.39

El modelo F, por tanto, colocaría d_1 en primer lugar mientras que la mayor parte de la gente consideraría más adecuado colocar d_2 primero.

Los operadores correspondientes a los modelos T de la página anterior poseen, por su parte, las siguientes propiedades:

- Cuando uno de sus operandos es 0 ó 1, el valor resultante es igual a uno de los dos operandos.

- En los demás casos, el valor resultante es mayor que el mayor de los operandos o menor que el menor de los operandos (compensación negativa)

La primera de las propiedades causa el problema comentado de los operadores max y min. La segunda propiedad, sin embargo, lo alivia. En el ejemplo de arriba, sustituyendo el mínimo por el producto (T_1), los valores asociados a los documentos d_1 y d_2 serían 0.16 y 0.39 respectivamente. Este resultado concuerda con el sentido común.

Sin embargo, el valor resultante en los modelos T, que es menor que el menor de los operandos o mayor que el mayor de ellos, ocasiona otro problema:

d = { (sistema,0.70), (recuperación,0.70), (información,0.70) }

q_1 = recuperación AND información
q_2 = sistema

T (d,q_1) < T (d,q_2) vg: T_1 (d,q_1) = 0.49
T_1 (d,q_2) = 0.70

De nuevo nos encontramos con otro resultado poco intuitivo: la similaridad del documento con la primera consulta es menor que con la segunda, algo con lo que mucha gente no estaría de acuerdo.

En cuanto a los operadores promedio, resaltemos que los operadores A_2 y A_4 son matemáticamente equivalentes en el caso binario:

$A_{2,AND}$ $\quad \gamma \max\{x,y\} + (1-\gamma)\min\{x,y\}$ $0 \le \gamma \le 0.5$

$A_{2,OR}$ $\quad \gamma \max\{x,y\} + (1-\gamma)\min\{x,y\}$ $0.5 \le \gamma \le 1$

Para hacer coincidir el rango de γ en los operadores $A_{2,AND}$ y $A_{2,OR}$, se sustituye γ por $(1-\gamma)$ en el primero de ellos:

$A_{2,AND}$ $\quad \gamma \min\{x,y\} + (1-\gamma)\max\{x,y\}$ $0.5 \le \gamma \le 1$

$A_{2,OR}$ $\quad \gamma \max\{x,y\} + (1-\gamma)\min\{x,y\}$ $0.5 \le \gamma \le 1$

Si sustituimos ahora γ por $(\gamma+1)/2$ obtenemos los operadores $A_{4,AND}$ y $A_{4,OR}$.

Por lo que respecta a A_1 y A_3, ambos manifiestan el problema de la compensación negativa de los operadores T para algunos casos. Además, A_1 siempre da 0 cuando uno de sus operandos es 0.

Por su parte, el resultado de A_4 (o, equivalentemente, A_2) siempre es mayor que el menor de los operandos y menor que el mayor de ellos, exceptuando, claro está, cuandos ambos operandos son iguales. Esta propiedad (compensación positiva) mejora la efectividad del sistema de recuperación de información.

Finalmente, el modelo booleano extendido E, no presenta el problema de la dependencia respecto a uno de los operandos como sucedía con los operadores T, A_1 y A_3. Además como A_4 (y A_2), el modelo E se caracterizan por obtener resultados incluidos siempre entre el mínimo y el máximo de sus operandos.

En resumen, los operadores parametrizados E y A_4 (A_2) resultan más adecuados para su utilización en sistemas de recuperación de información, ya que:

$$\min\{x,y\} \le A_{4.AND}(x,y) \le A_{4.OR}(x,y) \le \max\{x,y\}$$
$$\min\{x,y\} \le E_{AND}(x,y) \le E_{OR}(x,y) \le \max\{x,y\}$$

De hecho, las funciones definidas por ambos operadores son idempotentes (lo que implica que el resultado de las consultas “t”, “t AND t” y “t OR t” será el mismo), conmutativas (lo que indica que “a {AND|OR} b“ y “b {AND|OR} a” son equivalentes), estrictamente crecientes en cada uno de sus operandos (un incremento en uno de sus operandos incrementa el valor del resultado) y continuas.

Los resultados obtenidos con ambas parejas de operadores son similares, aunque la primera de ellas es más fácil de calcular. A_4 consigue una eficacia similar siendo más eficiente que E.

No obstante, dado que el número de posibles operadores para las operaciones AND y OR es infinito, nada nos permite excluir la posibilidad de que existan mejores operadores. De hecho, en recuperación de información se han propuesto otros operadores con buenas cualidades, como los operadores de Paice:

$$\frac{\sum_{i=1}^{n} r^{i-1} w_i}{\sum_{i=1}^{n} r^{i-1}} \qquad 0 \le r \le 1 \qquad \text{vg: } r=0.7$$

- P_{AND} con los pesos en orden ascendiente.
- P_{OR} con los pesos en orden descendiente.

Nótese que, en el caso binario, los operadores de Paice es equivalente matemáticamente a los operadores promedio A_4 (A_2):

$$P_{B.AND} = \frac{1}{1+r}\min\{x,y\} + \frac{r}{1+r}\max\{x,y\}$$
$$P_{B.OR} = \frac{1}{1+r}\max\{x,y\} + \frac{r}{1+r}\min\{x,y\}$$

La cardinalidad de los operadores

Los operadores T son funciones binarias, mientras que los operadores de Paice y las p-normas se han de definir como operadores n-arios al no ser asociativos. En el caso de las p-normas (modelo booleano extendido E), su expresión como operador n-ario es:

$$F(d, t_1\ AND\ \ldots\ AND\ t_n) = 1 - \sqrt[p]{\frac{(1-F(d,t_1))^p + \ldots + (1-F(d,t_n))^p}{n}}$$
$$F(d, t_1\ OR\ \ldots\ OR\ t_n) = \sqrt[p]{\frac{F(d,t_1)^p + \ldots + F(d,t_n)^p}{n}}$$

Los operadores promedio A tampoco son asociativos. Su versión n-aria viene dada por:

A_1 $\qquad (1-(1-w_1)\cdot\ldots\cdot(1-w_n))^{\gamma} \cdot (w_1\cdot\ldots\cdot w_n)^{1-\gamma}$

A_2 $\qquad \gamma \max\{w_1, w_2 \ldots w_n\} + (1-\gamma)\min\{w_1, w_2 \ldots w_n\}$

A_3 $\qquad \gamma \cdot (1-(1-w_1)\cdot\ldots\cdot(1-w_n)) + (1-\gamma)\cdot(w_1\cdot\ldots\cdot w_n)$

$A_{4,AND}$ $\gamma \min\{w_1 \cdot \ldots \cdot w_n\} + (1-\gamma)\frac{w_1 + \ldots + w_n}{n}$

$A_{4,OR}$ $\gamma \max\{w_1 \cdot \ldots \cdot w_n\} + (1-\gamma)\frac{w_1 + \ldots + w_n}{n}$

Desgraciadamente, todos los operadores que binarios que poseen las cualidades deseables para un sistema de recuperación de información (compensación positiva y no dependencia respecto a un solo operador) no pueden ser asociativos por definición:

$x < y \Rightarrow \theta(x,y) > x$ Compensación positiva

$\theta(\theta(x,y),y) > \theta(x,y)$ Monotonía

Si θ fuese asociativo, $\theta(\theta(x,y),y)=\theta(x,\theta(y,y))=\theta(x,y)$ por ser θ idempotente

Los operadores n-arios son necesarios para combinar términos que formen expresiones (AND) o especificar relaciones de sinonimia (OR). Cuando un usuario utiliza una expresión del tipo "a AND b AND c" o "a OR b OR c" espera que el resultado no dependa del orden de evaluación de los términos, por lo cual los operadores no asociativos han de definirse como funciones n-arias.

Sin embargo, como operadores n-arios, los operadores A_2, A_4 y P generan ordenaciones incorrectas en determinados casos:

✗ El operador n-ario A_2 (modelo de Waller-Kraft) considera sólo el máximo y el mínimo al calcular la similaridad entre una consulta y un documento (problema análogo al de los operadores T en el caso binario).

✗ El operador n-ario A_4 (modelo "infinity-one") no presenta el problema anterior al tener en cuanta todos sus operandos, pero sin embargo no les otorga la misma importancia a todos (el máximo y el mínimo influyen más).

✗ El operador de Paice, como el A_4, le da distinta importancia a sus distintos operandos.

✓ Las p-normas (modelo E) sí le dan la misma importancia a todos sus operandos. De todos los operadores expuestos son, a priori, los que presentan mejores propiedades matemáticas para obtener una alta efectividad.

De cualquier manera, aún se deben mejorar los modelos expuestos para tratar más adecuadamente las relaciones de proximidad, sinonimia e importancia relativa entre los términos de un documento.

Referencias bibliográficas

RAKESH AGRAWAL, ROBERTO BAYARDO & RAMAKRISHNAN SKIRANT
Athena: Text-based Interactive Management of Text Databases
IBM Quest Project

HSINCHUM CHEN & VASANT DHAR
A Knowledge-Based Approach to the Design of Document-Based Retrieval Systems
ACM, 1990

SUSAN T. DUMAIS
Tightly Coupling Search and Structure
SIGIR Workshop on Information Retrieval, 1997

C. LEE GILES, KURT D. BOLLACKER & STEVE LAWRENCE
CiteSeer: An Automatic Citation Indexing System
ACM Digital Libraries, 1998

STEVE LAWRENCE, KURT D. BOLLACKER AND C. LEE GILES
Indexing and Retrieval of Scientific Literature
ACM CIKM, 1999

JOON HO LEE, WON YONG KIM, MYOUNG HO KIM & YOON JOON LEE
On the Evaluation of Boolean Operators in the Extended Boolean Retrieval Framework
ACM SIGIR Conference on Research and Development in Information Retrieval, 1993

M. CATHERINE MCCABE, ABDUR CHOWDHURY, DAVID A. GROSSMAN & OPHIR FRIEDER
A Unified Environment for Fusion of Information Retrieval Approaches
ACM CIKM, 1999

ELLEN RILOFF & LEEAND LEE HOLLAAR
Text Databases and Information Retrieval
ACM Computing Surveys, Vol. 28, No. 1, March 1996

RAY SMITH
An Architecture for Textual Information Retrieval
TRW, 1988

S.K.M. WONG & W. ZIARKO
A Machine Learning Approach to Information Retrieval
ACM Conference on Research and Development in Information Retrieval, 1986

C.T. YU, W. MENG & S. PARK
A Framework for Effective Retrieval
ACM Transactions on Database Systems, Vol. 14, No. 2, June 1989