El modelo probabilístico: características y modelos derivados

Jesús Vilares
Departamento de Computación
Universidade da Coruña
Campus de Elviña
15071 – A Coruña (Spain)
jvilares@udc.es

Resumen

Presentamos en este trabajo una revisión del estado del arte de la familia de los modelos probabilísticos de recuperación de información. Partiendo de los principios básicos que sustentan estos modelos, estudiaremos diferentes modelos concretos: el modelo de independencia binaria —el más básico —, el ya clásico BM25 y, finalmente, los modelos DFR —uno de los últimos desarrollos.

1. Introducción

El planteamiento inicial del *modelo probabilístico* —si bien deberíamos hablar, más propiamente, de la *familia de los modelos probabilísticos*, en plural—, no es diferente del de otros modelos clásicos de *Recuperación de Información (RI)* [5, 12, 13, 8]. Como en cualquier sistema de RI, se parte de una *necesidad de información*, planteada por el usuario en forma de *consulta*, y de una *colección de documentos* sobre la cual realizar la búsqueda. A partir de las representaciones internas de ambos, el sistema intentará identificar aquellos documentos que satisfacen la necesidad, es decir, qué documentos son *relevantes* para la consulta.

La diferencia estriba en cómo y en base a qué dicha correspondencia entre consultas y documentos es calculada. En el caso del modelo vectorial [5, 12, 13, 8], por ejemplo, ésta es calculada empleando una base matemática formal, el álgebra vectorial, pero no existe ningún resultado teórico que nos permita afirmar que la forma en que calculamos dichas correspondencias es la correcta o la más adecuada. Aunque hacemos uso de una base matemática, en cierto modo estamos tanteando a ciegas, introduciendo cambios y aproximaciones en nuestro esquema de cálculo de correspondencias para luego comprobar experimentalmente si nuestras suposiciones eran correctas y permiten mejorar los resultados obtenidos.

Parece, pues, que no estamos empleando las herramientas adecuadas [13]. Dada una consulta, un sistema de RI tiene una comprensión incierta de la necesidad de información que dicha consulta representa. Del mismo modo, dadas las representaciones de la consulta y los documentos, el sistema sólo puede conjeturar, con un nivel de incertidumbre, si el contenido de un documento responde o no a la necesidad de información planteada. Sin embargo, la *teoría de probabilidades* nos dota de un marco formal que nos permite trabajar razonadamente en dicho ámbito. De este modo, las funciones de correspondencia desarrolladas estimarán la probabilidad de que un documento sea relevante para la consulta, en contraposición al concepto de grado o medida de relevancia de otros modelos [17]. En consecuencia, los documentos devueltos serán ordenados en base a sus probabilidades de relevancia estimadas respecto a la consulta [18], en lugar de en base a una medida de similaridad [24].

Planteadas inicialmente en los 70, las aproximaciones a RI basadas en modelos probabilísticos resurgieron con fuerza en los 90, gozando actualmente de gran atención por parte de la comunidad investigadora. A lo largo de este trabajo presentaremos este modelo, desde los principios teóricos que lo sustentan hasta las vías de investigación actuales. Primeramente, en la Sección 2 repasaremos algunos conceptos básicos de teoría de probabilidades. Seguidamente, en la Sección 3 introduciremos el principio de ordenación por probabilidad, sobre el cual se asienta el modelo. A continuación, en la Sección 4, describiremos el modelo de independencia binaria, uno de los primeros modelos probabilísticos y, también, uno de los más sencillos. La Sección 5 presenta una de sus evoluciones más conocidas y de mayor éxito, el Okapi BM25. A continuación, en la Sección 6, describimos los modelos DFR, entre los modelos probabilísticos más recientes y de mejores resultados. Finalmente, la Sección 7 cerrará este trabajo.

2. Conceptos Básicos de Teoría de Probabilidades

Antes de continuar, es conveniente repasar algunos conceptos previos de *teoría de probabilidades* [14] que utilizaremos repetidamente a lo largo de este trabajo. Sean:

$P(A)$ la probabilidad de que un suceso A ocurra
$P(\overline{A})$ la probabilidad de que un suceso A no ocurra

tenemos que:

$$P(A) + P(\overline{A}) = 1 \tag{1}$$

Repasemos ahora el concepto de *probabilidad condicionada*. Sean:

$P(A|B)$ la probabilidad condicionada de que ocurra un suceso A dado B, es decir, la probabilidad de que suceda A si ocurre B
$P(\overline{A}|B)$ la probabilidad condicionada de que no ocurra un suceso A dado B, es decir, la probabilidad de que no suceda A si ocurre B

de nuevo tenemos que:

$$P(A|B) + P(\overline{A}|B) = 1 \tag{2}$$

Por otra parte, el *teorema de Bayes* nos dice:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{3}$$

permitiéndonos expresar $P(A|B)$ en términos de $P(B|A)$.

Sea $P(A, B) = P(A \cap B)$ la *probabilidad conjunta* de dos sucesos A y B. Esta probabilidad se define, según la *regla de la cadena*, como:

$$P(A, B) = P(A \cap B) = P(A|B) \times P(B) = P(B|A) \times P(A) \tag{4}$$

Diremos que dos sucesos A y B son *independientes* si y sólo si:

$$P(A, B) = P(A \cap B) = P(A) \times P(B) \tag{5}$$

es decir, cuando la probabilidad condicionada de A dado B es simplemente la probabilidad de A y viceversa:

$$\begin{array}{l} P(A|B) = P(A) \\ P(B|A) = P(B) \end{array} \tag{6}$$

La denominada *asunción de Bayes ingenuo* (*Naive Bayes assumption*), nos dice que dado un suceso A compuesto por varios sucesos K_i independientes entre sí ($A = \{K_i \in A\}$) ocurre que:

$$P(A|B) = \prod_{K_i \in A} P(K_i|B) \tag{7}$$

Por otra parte, la *regla de decisión de Bayes* (*Bayes decision rule*) nos dice que, ante varias posibilidades de elección, escojamos aquélla que minimice la probabilidad de error:

$$\text{Decide } A' \text{ si } P(A'|C) > P(A_i|C) \text{ para todo } A_i \neq A' \tag{8}$$

Finalmente, definimos la *razón odds* (*odds ratio*) de un suceso A como:

$$O(A) = \frac{P(A)}{P(\overline{A})} \tag{9}$$

3. Principio de Ordenación por Probabilidad

El *principio de ordenación por probabilidad* (*probability ranking principle*) [16] constituye la base teórica justificativa sobre la que se asientan los modelos probabilísticos. Este principio nos dice —y demuestra— que la recuperación óptima es aquélla en la que los documentos son devueltos ordenados en orden decreciente de acuerdo a su probabilidad de relevancia respecta a la consulta. De este modo, un sistema de RI basado en un modelo probabilístico nos devolverá los documentos ordenados de acuerdo a la probabilidad $P(R|d_j, q)$ de que un documento d_j pertenezca al conjunto R de documentos relevantes para una consulta q o, en otras palabras, la probabilidad de que un documento d_j sea relevante para una consulta q [26].

Por otra parte, si además de una ordenación óptima queremos también un conjunto resultado óptimo, podemos aplicar la *regla de decisión de Bayes* (véase la Ecuación 8), según la cual los documentos devueltos —aquéllos considerados *relevantes*— deberían ser aquellos documentos para los cuales la probabilidad $P(R|d_j, q)$ de ser relevantes para la consulta es mayor que la probabilidad $P(\overline{R}|d_j, q)$ de no ser relevantes para la consulta:

$$\text{Un documento } d_j \text{ es } \textit{relevante} \text{ para una consulta } q \text{ si y sólo si } P(R|d_j, q) > P(\overline{R}|d_j, q) \tag{10}$$

4. El Modelo de Independencia Binaria

El *modelo de independencia binaria* (*binary independence model*) [13, 8, 7] es el modelo de recuperación más sencillo de la familia de los modelos probabilísticos. Este modelo, además de basarse en el principio de ordenación por probabilidad, aplica también la denominada *hipótesis clúster* [27], la cual supone que los términos que componen los documentos están distribuidos de forma diferente en el conjunto de documentos relevantes y en el conjunto de documentos no relevantes. A mayores, el modelo asume algunas suposiciones a la hora de estimar $P(R|d_j, q)$:

- Por *binario* entendemos *booleano* en el sentido de que, dados un término t_i y un documento d_j, sólo tendremos en cuenta si dicho término aparece o no en el documento, no cuántas veces aparece en él. De este modo, un documento d_j, compuesto por un conjunto de términos D_j, será representado internamente como un vector[1] de la forma:

  $$\overrightarrow{d_j} = (w_{1j}, w_{2j}, \ldots, w_{Mj})$$

  donde w_{ij}=1 si $t_i \in D_j$ —es decir, si el término t_i aparece en el documento d_j—, y w_{ij}=0 si $t_i \notin D_j$ —es decir, si el término t_i no aparece en el documento d.

  Lógicamente, lo mismo es aplicable al caso de una consulta q, constituida por un conjunto de términos Q y representada por un vector:

  $$\overrightarrow{q} = (w_{1q}, w_{2q}, \ldots, w_{Mq})$$

[1] Solución también empleada por el *modelo vectorial* [5, 12, 13, 8].

- Por *independencia* hacemos referencia al hecho de que el modelo supone que la distribución de un término t_i en la colección es independiente de la distribución de cualquier otro término t_j en la colección.[2]
- La relevancia de un documento es independiente de la relevancia de otros documentos.

Por otra parte, debemos tener en cuenta que, a la hora de estimar la probabilidad $P(R|d_j, q)$ de que un documento d_j sea relevante para una consulta q, realmente lo haremos en base a sus vectores representación $\overrightarrow{d_j}$ y $\overrightarrow{q}$. Trabajaremos, pues, en base a $P(R|\overrightarrow{d_j}, \overrightarrow{q})$. A su vez, a la hora de derivar una fórmula para esta probabilidad, aplicaremos dos transformaciones frecuentemente utilizadas en este tipo de cálculos:

1. Emplear razones odds en lugar de probabilidades (véase Ecuación 9).
2. Aplicar el teorema de Bayes (véase Ecuación 3).

En primer lugar, pasamos de trabajar en base a la probabilidad $P(R|\overrightarrow{d_j}, \overrightarrow{q})$ de que un vector documento $\overrightarrow{d_j}$ sea relevante para un vector consulta $\overrightarrow{q}$, a trabajar en base a la razón odds $O(R|\overrightarrow{d_j}, \overrightarrow{q})$ de que un vector documento $\overrightarrow{d_j}$ sea relevante para un vector consulta $\overrightarrow{q}$:

$$O(R|\overrightarrow{d_j}, \overrightarrow{q}) = \frac{P(R|\overrightarrow{d_j}, \overrightarrow{q})}{P(\overline{R}|\overrightarrow{d_j}, \overrightarrow{q})} \tag{11}$$

Al aplicar Bayes sobre esta expresión obtenemos:

$$O(R|\overrightarrow{d_j}, \overrightarrow{q}) = \frac{P(R|\overrightarrow{q})}{P(\overline{R}|\overrightarrow{q})} \times \frac{P(\overrightarrow{d_j}|R, \overrightarrow{q})}{P(\overrightarrow{d_j}|\overline{R}, \overrightarrow{q})} = O(R|\overrightarrow{q}) \times \frac{P(\overrightarrow{d_j}|R, \overrightarrow{q})}{P(\overrightarrow{d_j}|\overline{R}, \overrightarrow{q})} \tag{12}$$

y al asumir independencia entre los términos índice y aplicar la *asunción de Bayes ingenuo* (véase Ecuación 7), ocurre que:

$$\frac{P(\overrightarrow{d_j}|R, \overrightarrow{q})}{P(\overrightarrow{d_j}|\overline{R}, \overrightarrow{q})} = \prod_{i=1}^{M} \frac{P(w_{ij}|R, \overrightarrow{q})}{P(w_{ij}|\overline{R}, \overrightarrow{q})} \tag{13}$$

con lo que se obtiene:

$$O(R|\overrightarrow{d_j}, \overrightarrow{q}) = O(R|\overrightarrow{q}) \times \prod_{i=1}^{M} \frac{P(w_{ij}|R, \overrightarrow{q})}{P(w_{ij}|\overline{R}, \overrightarrow{q})} \tag{14}$$

Gracias a la propiedad conmutativa podemos, además, agrupar aquellos operandos correspondientes a términos t_i que aparecen en el documento d_j ($t_i \in D_j$) —donde w_{ij}=1—, y aquéllos correspondientes a términos t_i que no aparecen en el documento d_j ($t_i \notin D_j$) —donde w_{ij}=0:

$$O(R|\overrightarrow{d_j}, \overrightarrow{q}) = O(R|\overrightarrow{q}) \times \prod_{t_i \in D_j} \frac{P(w_{ij} = 1|R, \overrightarrow{q})}{P(w_{ij} = 1|\overline{R}, \overrightarrow{q})} \times \prod_{t_i \notin D_j} \frac{P(w_{ij} = 0|R, \overrightarrow{q})}{P(w_{ij} = 0|\overline{R}, \overrightarrow{q})} \tag{15}$$

Llegados a este punto, creemos conveniente simplificar la notación. Sea $p_i = P(w_{ij} = 1|R, \overrightarrow{q})$ la probabilidad de que un término t_i aparezca en un documento d_j relevante para la consulta, y sea $u_i = P(w_{ij} = 1|\overline{R}, \overrightarrow{q})$ la probabilidad de que un término t_i aparezca en un documento d_j no relevante para la consulta, obtenemos la siguiente tabla de contingencia:

Documento d_j	**Relevante** (R)	**No relevante** ($\overline{R}$)
Término t_i **presente** ($t_i \in D_j$, $w_{ij} = 1$)	p_i	u_i
Término t_i **no presente** ($t_i \notin D_j$, $w_{ij} = 0$)	$1 - p_i$	$1 - u_i$

[2] Aunque incorrecta —las palabras *Hong* y *Kong*, por ejemplo, tienden a co-ocurrir—, esta suposición es prácticamente una constante en los modelos de recuperación, ya que simplifica enormemente el modelo.

Haciendo las correspondientes sustituciones, obtenemos:

$$O(R|\overrightarrow{d_j}, \overrightarrow{q}) = O(R|\overrightarrow{q}) \times \prod_{t_i \in D_j} \frac{p_i}{u_i} \times \prod_{t_i \notin D_j} \frac{1-p_i}{1-u_i} \tag{16}$$

A mayores, dado que los términos que nos interesan son aquéllos que aparecen en la consulta ($t_i \in Q$), sería bueno eliminar los operandos correspondientes a aquellos términos que no aparecen en la consulta ($t_i \notin Q$). Para ello basta suponer que para esos términos la probabilidad de aparecer en un documento relevante o en uno no relevante es la misma, es decir $p_i = u_i$, con lo que numerador y denominador se simplifican:

$$\frac{p_i}{u_i} = \frac{p_i}{p_i} = 1 \qquad \frac{1-p_i}{1-u_i} = \frac{1-p_i}{1-p_i} = 1$$

obteniendo:

$$O(R|\overrightarrow{d_j}, \overrightarrow{q}) = O(R|\overrightarrow{q}) \times \prod_{\substack{t_i \in Q \\ t_i \in D_j}} \frac{p_i}{u_i} \times \prod_{\substack{t_i \in Q \\ t_i \notin D_j}} \frac{1-p_i}{1-u_i} \tag{17}$$

donde el producto de la izquierda hace referencia a los términos presentes en ambos, consulta y documento, y el producto de la derecha hace referencia a los términos presentes únicamente en la consulta, no en el documento. Por otra parte, si introducimos en una expresión un valor multiplicando y dividiendo simultáneamente, el valor de la expresión no varía. De igual modo, reordenando y reagrupando los factores —propiedades conmutativa y asociativa—, el valor de la expresión tampoco varía. Aplicando esto obtenemos:

$$\begin{aligned}
O(R|\overrightarrow{d_j}, \overrightarrow{q}) &= O(R|\overrightarrow{q}) \times \prod_{\substack{t_i \in Q \\ t_i \in D_j}} \tfrac{p_i}{u_i} \times \prod_{\substack{t_i \in Q \\ t_i \notin D_j}} \tfrac{1-p_i}{1-u_i} \\
&= O(R|\overrightarrow{q}) \times \prod_{\substack{t_i \in Q \\ t_i \in D_j}} \tfrac{p_i}{u_i} \times \prod_{\substack{t_i \in Q \\ t_i \notin D_j}} \tfrac{1-p_i}{1-u_i} \times \left(\prod_{\substack{t_i \in Q \\ t_i \in D_j}} \tfrac{1-p_i}{1-u_i} \times \prod_{\substack{t_i \in Q \\ t_i \in D_j}} \tfrac{1-u_i}{1-p_i} \right) \\
&= O(R|\overrightarrow{q}) \times \left(\prod_{\substack{t_i \in Q \\ t_i \in D_j}} \tfrac{p_i}{u_i} \times \prod_{\substack{t_i \in Q \\ t_i \in D_j}} \tfrac{1-u_i}{1-p_i} \right) \times \left(\prod_{\substack{t_i \in Q \\ t_i \notin D_j}} \tfrac{1-p_i}{1-u_i} \times \prod_{\substack{t_i \in Q \\ t_i \in D_j}} \tfrac{1-p_i}{1-u_i} \right) \\
&= O(R|\overrightarrow{q}) \times \left(\prod_{\substack{t_i \in Q \\ t_i \in D_j}} \tfrac{p_i}{u_i} \times \prod_{\substack{t_i \in Q \\ t_i \in D_j}} \tfrac{1-u_i}{1-p_i} \right) \times \left(\prod_{t_i \in Q} \tfrac{1-p_i}{1-u_i} \right) \\
&= O(R|\overrightarrow{q}) \times \left(\prod_{\substack{t_i \in Q \\ t_i \in D_j}} \tfrac{p_i \times (1-u_i)}{u_i \times (1-p_i)} \right) \times \left(\prod_{t_i \in Q} \tfrac{1-p_i}{1-u_i} \right)
\end{aligned} \tag{18}$$

A la hora de aplicar esta fórmula, debemos también tener en cuenta que realmente no estamos interesados en el valor $O(R|\overrightarrow{d_j}, \overrightarrow{q})$ en sí, sino en la ordenación de los documentos a la que da lugar. Por lo tanto, podemos simplificar la expresión anterior de tal forma que, si bien los valores obtenidos difieren, la ordenación se mantiene. Para ello eliminaremos de la expresión la razón odds inicial —ya que es constante para una consulta dada— y el producto final —ya que está definido para todos los términos de la consulta ($t_i \in Q$) y, por lo tanto, también es constante para una consulta dada. Si, además, aplicamos logaritmos —el logaritmo es una función monótona, por lo que la ordenación se mantiene—, el valor resultante, denominado RSV —de *Retrieval Status Value*—, y que será el que empleemos a la hora de ordenar los documentos devueltos, se define como:

$$RSV_{d_j q} = \log \prod_{\substack{t_i \in Q \\ t_i \in D_j}} \frac{p_i \times (1-u_i)}{u_i \times (1-p_i)} = \sum_{\substack{t_i \in Q \\ t_i \in D_j}} \log \frac{p_i \times (1-u_i)}{u_i \times (1-p_i)} \tag{19}$$

Finalmente, si consideramos por separado cada término de la consulta, obtenemos:

$$RSV_{d_j q} = \sum_{\substack{t_i \in Q \\ t_i \in D_j}} c_i \quad \text{con} \quad c_i = \log \frac{p_i \times (1-u_i)}{u_i \times (1-p_i)} = \log \frac{p_i}{1-p_i} + \log \frac{1-u_i}{u_i} = \log \frac{p_i}{1-p_i} - \log \frac{u_i}{1-u_i} = \log \frac{p_i \,/\, (1-p_i)}{u_i \,/\, (1-u_i)} \tag{20}$$

Como podemos ver, los c_i vienen dados por dos razones odds asociadas a los términos de la consulta. Se trata del logaritmo del cociente entre, por una parte, la razón odds de que el término de la consulta aparezca en un documento relevante ($p_i \,/\, (1-p_i)$), y por otra, la razón odds de que el término de la consulta aparezca en un documento no relevante ($u_i \,/\, (1-u_i)$). Por lo tanto, si un término de la consulta tiene la misma probabilidad de aparecer en un documento relevante que la de aparecer en un no relevante (i.e. $p_i = u_i$), el cociente será 1 y su logaritmo será 0, con lo que obtendríamos $c_i = 0$. Por otra parte, si la probabilidad de aparecer en un documento relevante es mayor que la de aparecer en un no relevante (i.e. $p_i > u_i$), el numerador será mayor que el denominador, el cociente será mayor que 1, y su logaritmo mayor que 0, con lo que obtendríamos $c_i > 0$. Finalmente, por el contrario, si la probabilidad de aparecer en un documento relevante es menor que la de aparecer en un no relevante (i.e. $p_i < u_i$), el numerador será menor que el denominador, el cociente será menor que 1, y su logaritmo menor que 0, con lo que obtendríamos $c_i < 0$.[3]

La cuestión, ahora, es cómo estimar c_i para una determinada consulta y una determinada colección de documentos.

4.1. Estimación de c_i

Llegados a este punto, dado que no se conoce el conjunto R de documentos relevantes, se hace necesario estimar los parámetros p_i y u_i para así poder calcular c_i.

En el caso de disponer de información sobre la relevancia de algunos documentos —obtenida mediante *realimentación* (*relevance feedback*) [5, 9]—, los parámetros pueden estimarse fácilmente [5, 7, 13]. Supongamos, pues, que el sistema ha devuelto un conjunto inicial de documentos para la consulta y que el usuario ha examinado algunos identificando cuáles de ellos son relevantes y cuáles no son relevantes.[4] Sea V el subconjunto de los documentos inicialmente devueltos que ha sido considerado relevante, y sea V_i el subconjunto de V cuyos documentos contienen el término t_i de la consulta. Lo que haremos será aproximar p_i mediante la distribución del término t_i en V:

$$p_i \approx \frac{|V_i|}{|V|} \tag{21}$$

donde $|V|$ y $|V_i|$ representan el número de elementos en los conjuntos V y V_i, respectivamente. De forma similar, y suponiendo que el resto de los documentos son no relevantes, aproximaremos u_i mediante:

$$u_i \approx \frac{df_i - |V_i|}{N - |V|} \tag{22}$$

donde N es el número de documentos que componen la colección y df_i es el número de documentos de la colección que contienen el término t_i. Para evitar problemas con valores pequeños de $|V|$ y $|V_i|$, se introducen, además, unos factores de ajuste, obteniendo finalmente:

$$p_i \approx \frac{|V_i| + 0.5}{|V| + 1} \tag{23}$$

$$u_i \approx \frac{df_i - |V_i| + 0.5}{N - |V| + 1} \tag{24}$$

[3] Desde el punto de vista operativo, y salvando las distancias, c_i es aquí asimilable al concepto clásico de *peso de un término* [5, 12, 13, 8].

[4] Otra posibilidad es emplear *realimentación automática*, también llamada *realimentación ciega* (*blind relevance feedback*) [12], consistente en asumir, sin necesidad de examinarlos, que los n primeros documentos devueltos son relevantes.

Finalmente, sustituyendo dichas estimaciones en la expresión de c_i de la Ecuación 20, y operando a continuación, se obtiene la expresión:

$$c_i \approx \log \frac{(|V_i| + 0.5) \,/\, (|V| - |V_i| + 0.5)}{(df_i - |V_i| + 0.5) \,/\, (N - df_i - |V| + |V_i| + 0.5)} \tag{25}$$

denominada también *peso Robertson-Sparck Jones* [18], y que tiene gran importancia en los esquemas de peso probabilísticos.

Pero, ¿qué ocurre si no podemos o no queremos emplear realimentación?. En ese caso se deberán emplear estimadores que no usen este tipo de información, lo cual complica notablemente su obtención. Dada su complejidad, hemos decidido no ir más allá en este trabajo, ya que creemos que se escapa al objetivo perseguido. Sin embargo, el lector podrá encontrar más información al respecto en [13].

5. El Okapi BM25

Una de las limitaciones del *modelo de independia binaria* descrito en la Sección 4, es que éste fue diseñado para ser empleado sobre textos cortos y de longitud más o menos similar —por ejemplo, repositorios de resúmenes. Es por ello que el modelo no presta atención a factores como la frecuencia de los términos dentro del documento —es binario, sólo tiene en cuenta la presencia o no del término— y la longitud del mismo. Por lo tanto, si queremos trabajar con otro tipo de colecciones más generales, se hace necesario contar con un modelo probabilístico más potente que sí tenga en cuenta esos factores.[5] Esta necesidad fue la que dió lugar al *BM25* [22, 21], uno de los modelos probabilísticos de referencia y, hasta la aparición de los modelos DFR —que veremos en la Sección 6—, uno de los mejores.

También llamado a menudo *Okapi BM25* o simplemente *Okapi* en referencia al sistema de RI para el que fue desarrollado [23], el *BM25* es un modelo probabilístico de recuperación perteneciente a la familia de *modelos Poisson-2* (*2-Poisson models*) [10]. Estos modelos asumen que las apariciones de un término en un documento tienen una naturaleza aleatoria, de tal forma que un documento es visto como una secuencia aleatoria de términos. Dicha distribución puede aproximarse mediante una *distribución de Poisson*, pero además asume que dicha distribución es diferente en aquellos documentos que tratan sobre el tema de ese término —llamados documentos *élite*—, y aquéllos que no tratan sobre el tema del término —llamados *no-élite*—, por lo que han de considerarse dos distribuciones de Poisson diferentes, de ahí la denominación *Poisson-2*. En su trabajo original [10], Harter empleaba el modelo únicamente para seleccionar términos de indexación, pero no les asociaba peso o medida de relevancia alguno. Serían trabajos posteriores, como el de Robertson [18], los que lo introdujesen.

No entraremos a detallar en este apartado los planteamientos teóricos detrás del modelo BM25,[6] sino que nos limitaremos a esbozar los sucesivos pasos mediante los cuales llegamos a la expresión final del modelo [8, 13]. Para ello partiremos inicialmente de la expresión de la Ecuación 20.

En primer lugar, nos centraremos en cómo tener en cuenta la *frecuencia del término en el documento* o *term frequency* (tf_{ij}), es decir, el número de veces que un término i aparece en un documento d_j. Debido a que la aplicación estricta de las distribuciones de Poisson daría lugar a expresiones excesivamente complejas, Robertson optó por aproximar dichas funciones por otras de comportamiento y forma similar, pero más simples y prácticas de calcular. De este modo, Robertson introdujo un nuevo factor corrector en el producto, una función de peso del término t_i en el documento d_j en base a su frecuencia tf_{ij}:

$$RSV_{d_j q} = \sum_{\substack{t_i \in Q \\ t_i \in D_j}} c_i \times \frac{(k_1 + 1) \times tf_{ij}}{k_1 + tf_{ij}} \tag{26}$$

[5] Factores que sí son tenidos en cuenta por otros modelos [5, 12, 13, 8], ya que esa información permite mejorar los resultados y aumentar la aplicabilidad del modelo.

[6] Si lo desea, el lector puede referirse a [23, 11, 8].

Como podemos ver, esta función aproximativa contiene una constante de ajuste k_1 cuyo valor controla la forma de la función, permitiéndonos ajustar el comportamiento de la misma: un valor k_1=0 devuelve al sistema su comportamiento binario original respecto a la frecuencia del término —es decir, no se tendría en cuenta tf_{ij}—, mientras que para valores altos de k_1 la función devolvería valores próximos a tf_{ij}. Por otra parte, del mismo modo que se ha tenido en cuenta la frecuencia del término en el documento, también se puede tener en cuenta la frecuencia del término en la consulta (tf_{iq}) introduciendo en el producto un factor análogo al anterior —al igual que en el caso anterior, existe una constante de ajuste (k_3). De este modo se obtiene:

$$RSV_{d_j q} = \sum_{\substack{t_i \in Q \\ t_i \in D_j}} c_i \times \frac{(k_1+1) \times tf_{ij}}{k_1 + tf_{ij}} \times \frac{(k_3+1) \times tf_{iq}}{k_3 + tf_{iq}} \tag{27}$$

La siguiente modificación propuesta por Robertson estuvo encaminada a tener en cuenta la longitud del documento. La razón para ello es que un documento muy largo tiene una probabilidad mucho mayor de que haya correspondencias con términos de la consulta, incluso aunque no sea relevante, ya que contiene muchas más palabras que un documento corto, pudiendo producirse más fácilmente correspondencias meramente casuales [25]. Para ello Robertson tuvo en cuenta dos hipótesis respecto a por qué un documento es más largo que otro [19]:

1. *Hipótesis del ámbito* (*scope hypothesis*): el documento más largo trata, a mayores, otros temas.

2. *Hipótesis de la verbosidad* (*verbosity hypothesis*): el documento más largo trata el mismo tema, pero en más detalle o bien, simplemente, con más "palabrería".

De este modo, Robertson introduce en la expresión la longitud dl_j del documento d_j actualmente considerado, multiplicando esta longitud por la constante k_1, ya que asume que la frecuencia de un término en un documento aumenta proporcionalmente a la longitud del mismo. A continuación, normaliza k_1 respecto a la longitud media de los documentos de la colección (dl_{avg}) dividiéndola por este valor. La razón para ello es que supone que k_1 ha sido fijada en base a dicha longitud media de los documentos de la colección. Finalmente, introduce un último factor de ajuste b, con valor entre 0 y 1, el cual permite al usuario controlar en qué medida la longitud del documento es tenida en cuenta. De este modo, un valor extremo b=0 "desactivaría" las modificaciones encaminadas a tener en cuenta la longitud, mientras que valores mayores irían introduciendo el factor modificador cada vez en mayor medida hasta aplicarlo totalmente en el caso de b=1. La expresión final resultante es:

$$RSV_{d_j q} = \sum_{\substack{t_i \in Q \\ t_i \in D_j}} c_i \times \frac{(k_1+1) \times tf_{ij}}{K + tf_{ij}} \times \frac{(k_3+1) \times tf_{iq}}{k_3 + tf_{iq}} \tag{28}$$

donde c_i es el factor inicial, tal como se define en la Ecuación 20
K se calcula como $K = k_1 \times ((1-b) + b \times dl_j / dl_{avg})$
k_1, k_3 y b son parámetros constantes de ajuste
tf_{ij} es la frecuencia del término t_i en el documento d_j
tf_{iq} es la frecuencia del término t_i en la consulta q
dl_j es la longitud del documento d_j
dl_{avg} es la longitud media de los documentos de la colección

Idealmente, los valores de los parámetros k_1, k_3 y b deberían ser ajustados experimentalmente en base a la colección y al tipo de consultas empleadas, si bien en la práctica suelen emplearse ciertos valores por defecto [20, 8, 13]: k_1 y k_3 entre 1.2 y 2 —normalmente 1.2—, pudiendo aumentar k_3 a valores entre 7 y 1000 (virtualmente infinito) en el caso de consultas largas; y b=0.75, aunque valores más pequeños de b a veces reportan mejoras.

6. El Paradigma DFR

Al hablar del paradigma DFR [4, 1] —de *Divergence From Randomness* o *divergencia respecto a la aleatoriedad*—, y cuyo principal exponente es el sistema TERRIER [2], no deberíamos tanto hablar de un modelo de recuperación sino de una *metodología* para construir modelos de recuperación.

Al igual que en el caso del BM25 —descrito en la Sección 5—, los modelos DFR parten del modelo *Poisson-2* definido por Harter [10]. Sin embargo, los modelos DFR guardan importantes diferencias respecto a otros modelos probabilísticos:

- Los modelos DFR no están basados en el *principio de ordenación por probabilidad* —véase Sección 3—, ya que no trabajan en base al concepto de *probabilidad de relevancia* de un documento, sino en base a los conceptos de *contenido informativo* (*informative content*) y *ganancia de información* (*information gain*) [4]. De esta forma, los documentos no son devueltos de acuerdo a su probabilidad de relevancia respecto a la consulta, sino respecto a la ganancia de información obtenida al devolver dicho documento —reflejada en un valor RSV—, retomando además el concepto de *peso* w_{ij} *de un término* t_i *en un documento* d_j [5, 12, 13, 8] a la hora de calcular dicha ganancia a nivel de término:

$$RSV_{d_j q} = \sum_{t_i \in Q} w_{ij} \tag{29}$$

- Los modelos DFR son modelos *no paramétricos*, por lo que no es necesario ningún ajuste experimental de parámetros, como en el caso del BM25.

- Como ya hemos dicho, se trata de una *metodología* para construir modelos de recuperación, de tal forma que distintas configuraciones darán lugar a distintos modelos DFR.

La idea sobre la que se basan los modelos DFR es sencilla. Si asumimos que la distribución de los términos en los documentos a lo largo de la colección debería ser aleatoria —de acuerdo con el modelo *Poisson-2* [10]—, entonces podemos medir la cantidad de información portada por un término t_i en un documento d_j en base a la diferencia entre su distribución real en ese documento y su distribución esperada según el modelo aleatorio. De forma más sencilla, si una palabra aparece en un documento muchas más veces de lo que cabría de esperar, entonces parece lógico suponer que dicho documento trata ese tema.

A la hora de definir un modelo DFR debemos definir sus tres componentes, los cuales tienen una plasmación directa en la expresión de cálculo del peso w_{ij} de un término t_i en un documento d_j:

$$w_{ij} = tf_{iq} \times Inf_1(tfn_{ij}) \times P_{risk}(tfn_{ij}) \tag{30}$$

donde tf_{iq} es la frecuencia del término t_i en la consulta q

Inf_1 es el *contenido informativo* del término t_i en el documento d_j (ver Sección 6.1)

P_{risk} es una expresión del *riesgo* asumido al aceptar el término t_i como descriptor válido del documento d_j (ver Sección 6.2)

tfn_{ij} es la frecuencia tf_{ij} del término t_i en el documento d_j tras ser normalizada en base a la longitud del documento (ver Sección 6.3)

A continuación, describiremos en mayor detalle esos tres componentes.

6.1. Componente 1: Modelo Aleatorio

El primero de los componentes de un modelo DFR es el *modelo aleatorio* (*randomness model*) según el cual asumimos que están distribuidos los términos, y que vendrá dado por una función de probabilidad $Prob_1$ [4, 1], donde $Prob_1(tf_{ij})$ es la probabilidad de que el término t_i aparezca tf_{ij} veces en el documento d_j. Uno de los casos más sencillos es emplear una distribución binomial:

$$Prob_1(tf_{ij}) = \binom{TF_i}{tf_{ij}} \times p^{tf_{ij}} \times q^{TF_{ij} - tfij} \quad \text{con} \quad p = \frac{1}{N} \quad \text{y} \quad q = 1 - p \tag{31}$$

donde tf_{ij} es la frecuencia del término t_i en el documento d_j
TF_i es la frecuencia total del término t_i en la colección
N es el número de documentos en la colección

Otra posibilidad, algo más compleja, es emplear una distribución geométrica:

$$Prob_1(tf_{ij}) = -\log_2 \left(\left(\frac{1}{1+\lambda} \right) \times \left(\frac{\lambda}{1+\lambda} \right)^{tf_{ij}} \right) \quad \text{con} \quad \lambda = \frac{TF_i}{N} \tag{32}$$

Según la función $Prob_1$ que tomemos, obtendremos un modelo diferente. El lector podrá encontrar en [4, 1] más información al respecto.

Una vez definida $Prob_1$, ya se puede calcular el *contenido informativo* (*informative content*) de un término en un documento (Inf_1). Intuitivamente, podemos considerar que en la colección existen dos tipos de palabras [10]. Por una parte, estarían las denominadas *palabras "de especialidad"* (*specialty words*), aquéllas de mayor contenido informativo y que se concentran en los documentos *élite* —por lo tanto más útiles para el proceso de recuperación—, y las cuales suponemos que difieren en su distribución de lo esperado según el modelo aleatorio. Por otra parte, estarían las denominadas *palabras "de no-especialidad"* (*nonspecialty words*), de escaso contenido informativo —caso de las *stopwords*—, y para las cuales asumimos una distribución aleatoria a lo largo de la colección. De este modo, si, de acuerdo a nuestro modelo de distribución, un término tiene una alta probabilidad de aparecer en un documento, asumiremos que se trata de una palabra *"de no-especialidad"*, es decir, de escaso contenido informativo. Por el contrario, si un término tiene una probabilidad baja de aparecer en un documento, entonces estaríamos ante una palabra *"de especialidad"*, aquellas palabras que proveen a un documento de su contenido informativo. En base a esto, se define el *contenido informativo* (*informative content*) de un término en un documento (Inf_1) como:

$$Inf_1 = -\log_2 Prob_1 \tag{33}$$

6.2. Componente 2: Primera Normalización

El segundo componente de un modelo DFR es el denominado comúnmente *primera normalización* (*first normalization*) [4, 1].

La idea básica detrás de esta normalización es la siguiente. Si un término poco común —*"de especialidad"*— no aparece en un documento, podemos suponer que su probabilidad de ser informativo, en relación al tema que trata ese documento, es baja o nula. De este modo, si aceptamos dicho término como un descriptor válido de ese documento, estaremos asumiendo un *riesgo risk* muy alto [15], ya que existen indicadores que apuntan a que es poco fiable. Por el contrario, si un término poco común aparece repetidamente en un documento, entonces podemos suponer que su probabilidad de ser informativo, en relación al tema del documento, es muy alta. En consecuencia, si lo tomamos como descriptor, el riesgo asumido en este caso sería muy bajo.

Si denominamos $Prob_2$ a dicha probabilidad de un término t_i de ser informativo en relación al tema tratado en d_j, podemos definir la función de riesgo P_{risk} asociada a tomar t_i como término representativo de d_j como:

$$P_{risk} = 1 - Prob_2 \tag{34}$$

La normalización a la que hace referencia este componente consiste en multiplicar Inf_1 por P_{risk} —véase Ecuación 30—, y así ponderar el contenido informativo inicial Inf_1 de un término en un documento en base al *riesgo* asumido al tomar dicho término como un descriptor válido de ese documento.

Finalmente, en [4, 1] se describen dos modelos para la función P_{risk}, uno basado en la ley de sucesión de Laplace, y que es referido como *normalización L*:

$$P_{risk} = \frac{1}{tf_{ij} + 1} \tag{35}$$

y otro basado en distribuciones binomiales, y que es referido como *normalización B*:

$$P_{risk} = \frac{TF_i + 1}{df_i \times (tf_{ij} + 1)} \tag{36}$$

donde df_i es el número de documentos de la colección que contienen el término t_i.

6.3. Componente 3: Segunda Normalización

El tercer y último componente de un modelo DFR es el denominado *segunda normalización* (*second normalization*) [4, 1], y que persigue normalizar la frecuencia tf_{ij} del término t_i en el documento d_j en base a la longitud del documento y a la longitud media de un documento de la colección. El valor normalizado resultante, tfn_{ij}, es el valor empleado finalmente a la hora de calcular el peso, en lugar de la frecuencia tf_{ij} inicial —ver Ecuación 30. A este respecto, se describen dos normalizaciones en [4, 1]:

$$tfn_{ij} = tf_{ij} \times \frac{dl_{avg}}{dl_j} \tag{37}$$

$$tfn_{ij} = tf_{ij} \times \log_2 \left(1 + \frac{dl_{avg}}{dl_j}\right) \tag{38}$$

donde dl_j es la longitud del documento d_j
dl_{avg} es la longitud media de los documentos de la colección

7. Conclusión

A lo largo de este trabajo hemos hecho una revisión del estado del arte de la familia de los modelos probabilísticos de recuperación de información. Partiendo de los principios básicos que sustentan estos modelos, hemos descrito diferentes modelos concretos, desde el más básico —el modelo de independencia binaria—, hasta sus últimos desarrollos —los modelos DFR—, pasando por el ya clásico BM25. Hemos optado por dejar fuera, al considerarlas fuera del ámbito de este trabajo, otras aproximaciones de base probabilística como las *redes Bayesianas* (*Bayesian networks*) [8, 7, 5] o los *modelos de lenguaje* (*language modeling*) [15, 6, 13] —cuyo mejor exponente es el sistema Lemur [3]. El lector puede consultar la bibliografía al respecto si tiene interés.

Referencias

[1] Divergence From Randomness (DFR) Framework. `http://ir.dcs.gla.ac.uk/wiki/Terrier/LiveDoc/DFRDescription`.

[2] TERabyte RetrIEveR: Terrier Information Retrieval Platform. `http://ir.dcs.gla.ac.uk/terrier/`.

[3] The Lemur Toolkit. `http://www.lemurproject.org`.

[4] G. Amati and C.J. van Rijsbergen. Probabilistic models of Information Retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389, 2002.

[5] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley and ACM Press, Harlow, England, 1999.

[6] W.B. Croft and J. Lafferty, editors. *Language Modeling for Information Retrieval*, volume 13 of *The Information Retrieval Series*. Kluwer Academic Publishers, 2003.

[7] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255, 1992.

[8] E. Greengrass. Information retrieval: A survey. Technical Report TR-R52-008-001, United States Department of Defense, 2001.

[9] D. Harman. Relevance feedback revisited. In *Proceedings of the 15th annual international ACM SIGIR conference on research and development in Information Retrieval (SIGIR'92), June 21-24, Copenhagen, Denmark*, pages 1–10. ACM Press, 1992.

[10] S.P. Harter. A probabilistic approach to automatic keyword indexing, Parts I & II. *Journal of the American Society for Information Science*, 26(4):197–206, 280–289, 1975.

[11] K. Spark Jones, S. Walker, and S.E. Robertson. A probabilistic model of information retrieval: development and comparative experiments, Parts I & II. *Information Processing & Management*, 36(6):779–808, 809–840, 2000.

[12] G. Kowalski. *Information Retrieval Systems: Theory and Implementation*. The Kluwer international series on Information Retrieval. Kluwer Academic Publishers, Boston-Dordrecht-London, 1997.

[13] C.D. Manning, P. Raghavan, and H. Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England, 2008.

[14] C.D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge (Massachusetts) and London (England), 1999.

[15] J.M. Ponte and W.B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in Information Retrieval (SIGIR'98), August 24-28 1998, Melbourne, Australia*, pages 275–281. ACM Press, 1998.

[16] S.E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, (33):126–148, 1977.

[17] S.E. Robertson and N.J. Belkin. Ranking in principle. *Journal of Documentation*, 34(2):93–100, 1978.

[18] S.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Sciences*, (27):129–146, May–June 1976.

[19] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on research and development in Information Retrieval (SIGIR'94)*, pages 232–241. Springer-Verlag New York, Inc., 1994.

[20] S.E. Robertson and S. Walker. Microsoft Cambridge at TREC-9: Filtering track. In E.M. Voorhees and D.K. Harman, editors, *NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC 9)*, pages 361–368, Gaithersburg, MD, USA, 2001. Department of Commerce, National Institute of Standards and Technology.

[21] S.E. Robertson, S. Walker, M.M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In D. K. Harman, editor, *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)*, pages 73–96, Gaithersburg, MD, USA, 1996. Department of Commerce, National Institute of Standards and Technology.

[22] S.E. Robertson, S. Walker, K. Spark Jones, M.M. Beaulieu, and M. Gatford. Okapi at TREC-3. In D. K. Harman, editor, *NIST Special Publication 500-225: Overview of the Third Text REtrieval Conference (TREC 3)*, pages 109–126, Gaithersburg, MD, USA, 1995. Department of Commerce, National Institute of Standards and Technology.

[23] S. Robertson. *TREC: Experiment and Evaluation in Information Retrieval*, chapter How Okapi came to TREC, pages 287–299. MIT Press, 2005.

[24] G. Salton, editor. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[25] G. Salton and C. Buckley. Term-weighting approaches in automatic retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[26] C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2^{nd} edition, 1979.

[27] C.J. van Rijsbergen and K. Spark Jones. A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, 29(3):251–257, 1973.