El Modelo Probabilístico:

Características y Modelos Derivados

Jesús Vilares

Grupo de Lengua y Sociedad de la Información (LYS)

Universidade da Coruña

jvilares@udc.es

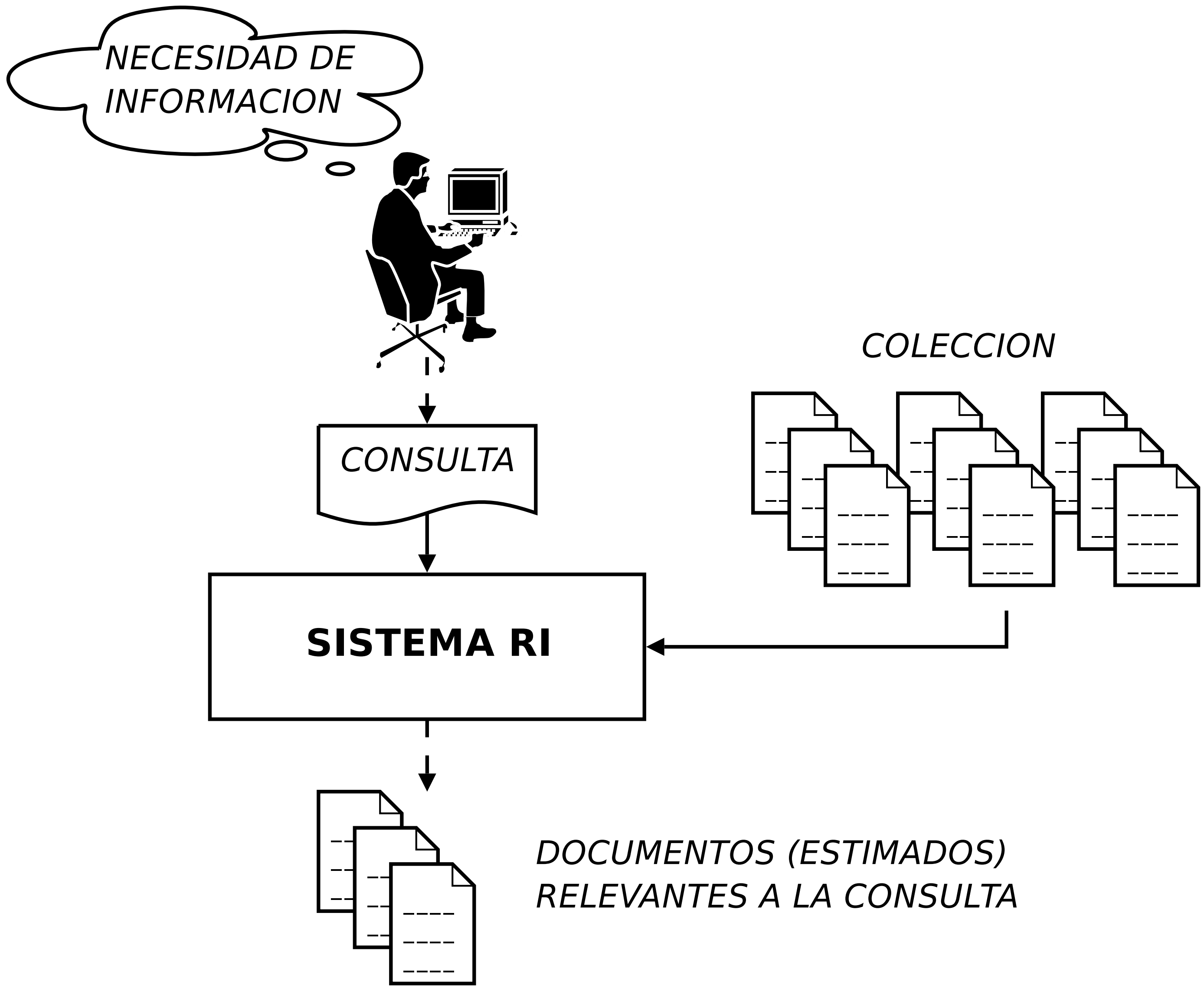
Índice

- Introducción
- Conceptos de Teoría de Probabilidades
- Principio de Ordenación por Probabilidad
- Modelo de Independencia Binaria
- Okapi BM25
- Paradigma DFR
- Conclusión

Índice

- **Introducción**
- Conceptos de Teoría de Probabilidades
- Principio de Ordenación por Probabilidad
- Modelo de Independencia Binaria
- Okapi BM25
- Paradigma DFR
- Conclusión

Recuperación de Información (RI)



Terminología

- **Documento:** unidad de texto almacenada y disponible para su recuperación; p.ej., páginas web, artículos de prensa, tesis, ...
- **Colección:** repositorio de documentos en los que buscar
- **Términos:** unidades léxicas (palabras) que componen un documento/consulta
- **Consulta (*query*)**: **representación** en forma de términos, de la necesidad de información del usuario

Terminología (cont.)

- **Relevancia de un documento**:
  - Calculada por el sistema respecto a la *consulta*
  - Juzgada por el usuario respecto a la **necesidad de información** en su cabeza (**subjetividad**)

- **Ordenación (*ranking*)**: los documentos suelen devolverse ordenados **por relevancia**

- **Peso de un término**: medida de su **representatividad**
  - Frecuencia dentro del documento
  - Distribución dentro de la colección
  - Longitud del documento

Paradigma *Bag-of-Terms*

- **Def.:** representación de documentos/consultas como **conjunto** de *términos índice*
- ***Ppo. de composicionalidad de Frege***: "la semántica de un objeto puede obtenerse a partir de la semántica de sus componentes"
  - Si una palabra aparece en un texto, dicho texto trata dicho tema
  - **Si una consulta y un documento comparten uno/más términos índice, el documento debería tratar el tema de la consulta**

Modelos de Recuperación

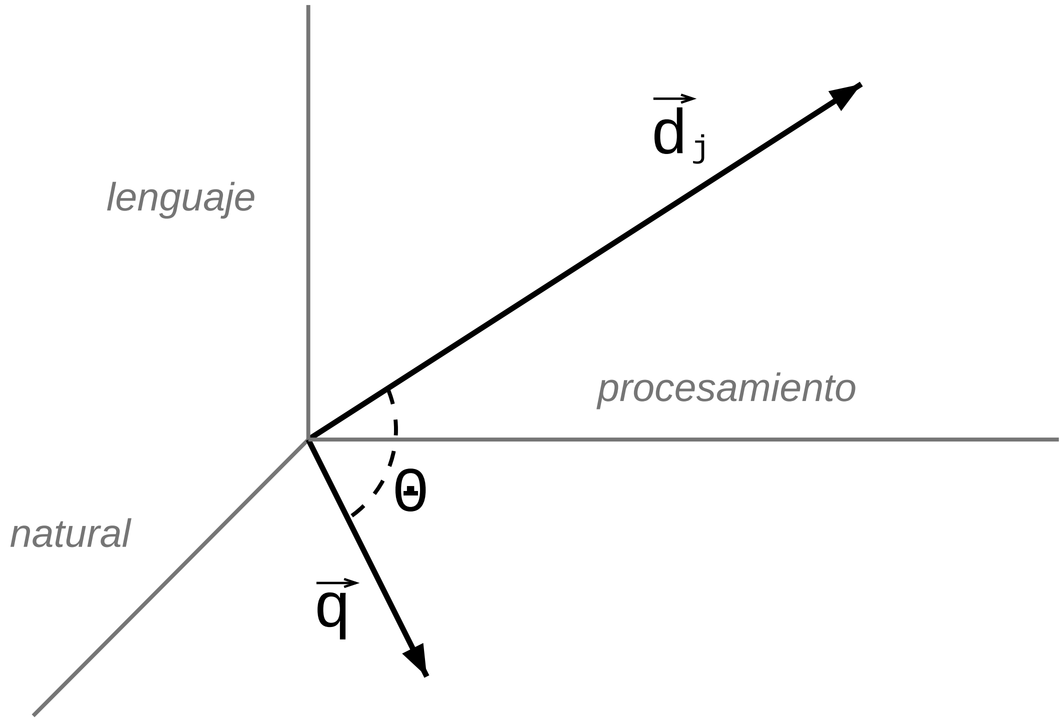
- Establecen:
  - Cómo representar los documentos
  - Cómo representar la consulta
  - Cómo compararlos

Otros Modelos

- **Modelo vectorial como ejemplo**
- Base matemática: **álgebra vectorial**
- **Consultas y documentos representados como vectores** en un espacio multidimensional
  - 1 dimensión por término vocabulario
  - P.ej. Vocabulario tamaño $M \rightarrow$ espacio M-dimensional
    - Documento d_j: vector $\overrightarrow{d_j} = (w_{1j}, w_{2j}, \ldots, w_{Mj})$
    - Consulta q: vector $\overrightarrow{q} = (w_{1q}, w_{2q}, \ldots, w_{Mq})$

  donde $w_{ij} \geq 0$ y $w_{iq} \geq 0$ los pesos del término t_i en d_j y q

Otros Modelos (cont.)



- Si los vectores de consulta y documento están próximos, asumimos que documento es similar a la consulta (i.e., posiblemente relevante)

Otros Modelos (cont.)

- **Medida proximidad (similaridad):** coseno del ángulo Θ formado por los vectores:

$$sim(d_j, q) = cos\,(\Theta) = \frac{\overrightarrow{d_j} \bullet \overrightarrow{q}}{|\overrightarrow{d_j}| \times |\overrightarrow{q}|} = \frac{\sum_{i=1}^{M} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{M} w^2{}_{ij}} \times \sqrt{\sum_{i=1}^{M} w^2{}_{iq}}}$$

Otros Modelos (cont.)

- **¿Base formal?**
  Sí.

- **Forma calcular correspondencias, ¿es la mejor/más adecuada?**
  No sabemos, no hay nada que nos lo permite afirmar.

(Familia) Modelos Probabilísticos

- **Sistema IR:**
  - Comprensión incierta de la necesidad/consulta.
  - Conjeturar acerca de si el contenido del documento es relevante.

- Marco formal de trabajo: **teoría de probabilidades**
  - **Probabilidad de relevancia** vs. medida similaridad

Índice

- Introducción
- **Conceptos de Teoría de Probabilidades**
- Principio de Ordenación por Probabilidad
- Modelo de Independencia Binaria
- Okapi BM25
- Paradigma DFR
- Conclusión

Conceptos de Teoría de Probabilidades

$P(A)$ probabilidad de que un suceso A ocurra

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

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

$P(A|B)$ probabilidad (condicionada) de que suceda A si ocurre B

$P(\overline{A}|B)$ probabilidad (condicionada) de que no suceda A si ocurre B

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

A y B **independientes entre sí**:

$$P(A|B) = P(A) \qquad P(B|A) = P(B)$$

$$P(A, B) = P(A \cap B) = P(A) \times P(B)$$

Conceptos de Teoría de Probabilidades (cont.)

- ***Teorema de Bayes***:

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

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

- ***Razón odds*** (*odds ratio*) de un suceso A:

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

Índice

Ppo. de Ordenación por Probabilidad

- **Base de los modelos probabilísticos**:

  ”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”

Ppo. de Ordenación por Probabilidad (cont.)

- Sean:

  $P(R|d_j, q)$ probabilidad de que un documento d_j sea relevante para una consulta q

  $P(\overline{R}|d_j, q)$ probabilidad de que un documento d_j no sea relevante para una consulta q

  - Documentos **devueltos por orden de probabilidad** de relevancia $P(R|d_j, q)$

  - Documento **es relevante** si $P(R|d_j, q) > P(\overline{R}|d_j, q)$

Índice

Bases del Modelo

- El más sencillo de los probabilísticos.
- ***Hipótesis clúster***:

  "los términos están distribuidos de forma diferente en los documentos relevantes y no relevantes"

- **Binario (booleano)**: sólo tendremos en cuenta si un término aparece o no en un documento, no cuántas veces:

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

donde $w_{ij} = 1$ si $t_i \in D_j$ (término t_i está en documento d_j)
$w_{ij} = 0$ si $t_i \notin D_j$ (término t_i no está en documento d_j)

Bases del Modelo (cont.)

- **Independencia:**
  - Distribución de un término en la colección **independiente** de la de otros
  - Relevancia de un documento **independiente** de la de otros

Formulación

- Trabajaremos con $O(R|\overrightarrow{d_j}, \overrightarrow{q})$ en lugar de con $P(R|\overrightarrow{d_j}, \overrightarrow{q})$:

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

- Al aplicar el **Teorema de Bayes**:

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

- Al asumir que **los términos son independientes** entre sí:

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

Formulación (cont.)

- Agrupamos los operandos de los términos **según aparezcan o no en el documento**:

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

- **Simplificamos la notación**:

$p_i = P(w_{ij} = 1|R, \overrightarrow{q})$ prob. término t_i aparezca en doc. relevante

$u_i = P(w_{ij} = 1|\overline{R}, \overrightarrow{q})$ prob. término t_i aparezca en doc. no relevante

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

Formulación (cont.)

- Obviamos **términos ajenos a la consulta**:

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

- Operando sucesivamente:

$$(\ldots)$$

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

Formulación (cont.)

- **Sólo nos interesa la ordenación**, no el valor concreto:
  - Eliminamos factores constantes (mantiene ordenación)
  - Aplicamos logaritmos (mantiene ordenación)
  - ***Retrieval Status Value***

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

- Considerando cada término de la consulta por separado:

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

Formulación (cont.)

$$c_i = \log \frac{p_i \,/\, (1 - p_i)}{u_i \,/\, (1 - u_i)}$$

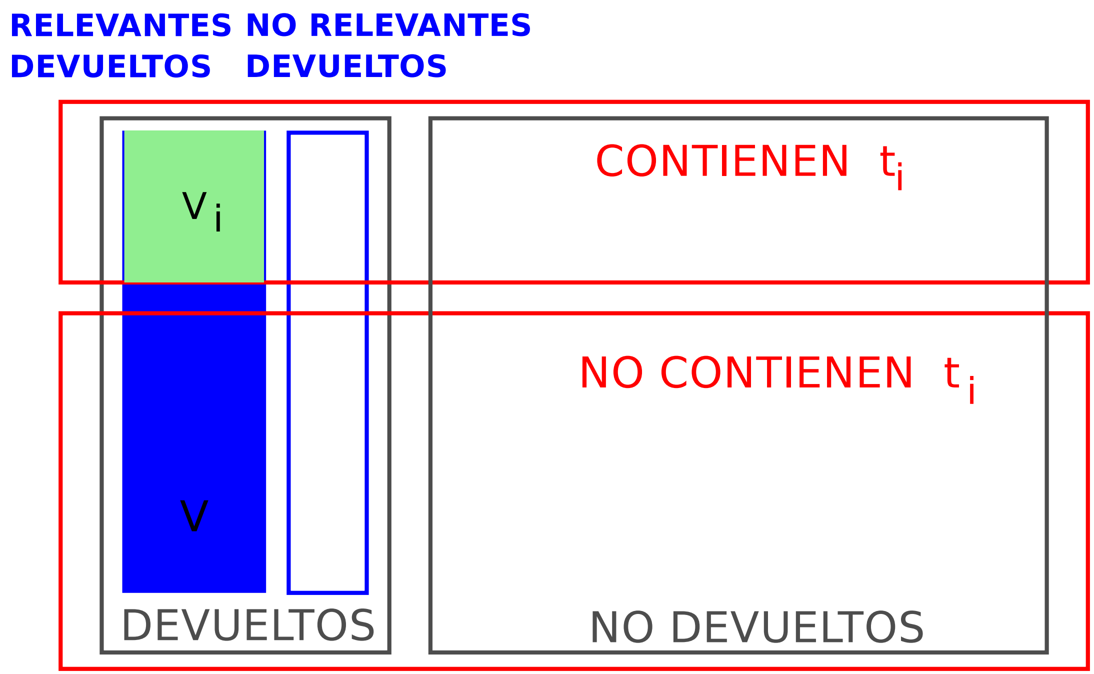
- Témino más probable en relevantes ($p_i > u_i$): $c_i > 0$.
- Témino más probable en no relevantes ($p_i < u_i$): $c_i < 0$.
- Término igualmente probable ($p_i = u_i$): $c_i = 0$.

Estimación de Probabilidades

$$c_i = \log \frac{p_i \,/\, (1 - p_i)}{u_i \,/\, (1 - u_i)}$$

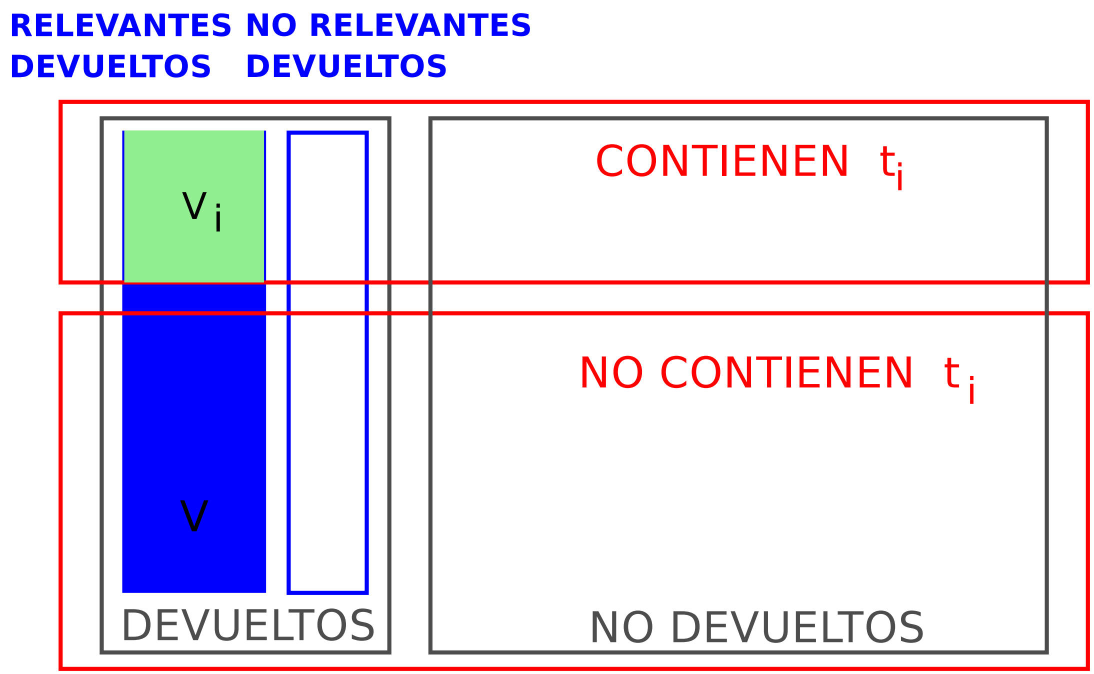
- **Problema**: desconocemos p_i y u_i
- **Solución**: **estimación** a partir de subconjunto resultado inicial (*relevance feedback*):
  - Obtenemos conjunto resultado inicial
  - Comprobamos cuáles son relevantes
  - Estimamos p_i y u_i a partir de estos conjuntos

Estimación de Probabilidades (cont.)



- Conocemos:
  - $|V|$, nº docs. relevantes devueltos
  - $|V_i|$, nº docs. relevantes devueltos contienen término t_i
  - N, nº docs. en colección
  - df_i, nº docs. en colección contienen término t_i

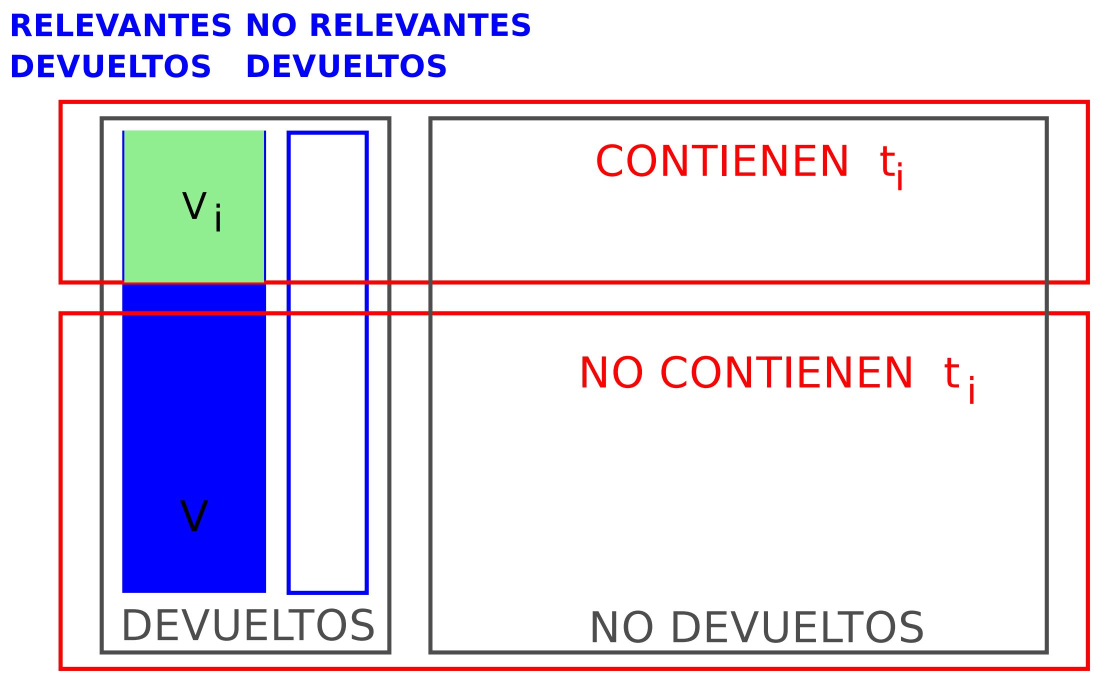
Estimación de Probabilidades (cont.)



- Aproximamos p_i mediante la **proporción de docs. relevantes devueltos que contienen término** t_i:

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

Estimación de Probabilidades (cont.)



- Suponiendo resto son *no relevantes*, aproximamos u_i mediante la **proporción de docs. no relevantes que contienen término** t_i:

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

Estimación de Probabilidades (cont.)

- Sustituyendo y operando:
  - Factores de ajuste

$$
\begin{aligned}
c_i &= \log \frac{p_i \,/\, (1-p_i)}{u_i \,/\, (1-u_i)} \\
&(\ldots) \\
&\approx \log \frac{(|V_i|+0{,}5) \,/\, (|V|-|V_i|+0{,}5)}{(df_i-|V_i|+0{,}5) \,/\, (N-df_i-|V|+|V_i|+0{,}5)}
\end{aligned}
$$

denominado ***peso Robertson-Sparck Jones***

Índice

Okapi BM25

- Modelo de referencia (entre los mejores)
- **Evolución** del *modelo de dependencia binaria*, introduce:
  - Nº apariciones del término en el documento
  - Longitud del documento

Formulación: Base Inicial

- Partimos de la expresión del *modelo de independencia binaria* básico:

$$RSV_{d_j q} = \sum_{\substack{t_i \in Q \\ t_i \in D_j}} c_i$$

Formulación: Frec. Término

- Ponderar nº apariciones del término en el documento: *frecuencia del término* t_i *en el documento* d_j (tf_{ij})
- Introducir función de peso del término en el documento en base a su frecuencia:

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

- Constante de ajuste k_1:
  - $k_1 = 0$: comportamiento binario original
  - k_1 muy alto: devolvería valores próximos a tf_{ij}

Formulación: Frec. Término (cont.)

- Ídem para frecuencia de los términos en la consulta:

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

Formulación: Longitud Doc.

- Ponderar longitud del documento
- Introducimos longitud dl_j del documento d_j, normalizada respecto a la longitud media de los documentos de la colección (dl_{avg}):

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

  con $K = k_1 \times ((1-b) + b \times dl_j / dl_{avg})$

- Constante de ajuste $b \in [0,1]$:
  - $b = 0$: se desestima longitud
  - $b = 1$: aplicación plena

Índice

- Introducción
- Conceptos de Teoría de Probabilidades
- Principio de Ordenación por Probabilidad
- Modelo de Independencia Binaria
- Okapi BM25
- **Paradigma DFR**
- Conclusión

Paradigma DFR

- *Divergence From Randomness* (DFR): *metodología* para construir modelos de recuperación

- **Diferencias** respecto modelos probabilísticos ”clásicos”:
  - *Metodología*, no modelo.
  - *No paramétrico*: no hay parámetros a ajustar (ej. k_1, k_3 y b en BM25).
  - *Ganancia de información* vs. probabilidad de relevancia.

- **Idea**:
  - Asumir distribución aleatoria de los términos en los docs.
  - Si una palabra aparece en un doc. mucho más de lo esperado, ese doc. trata ese tema.

Paradigma DFR: Componentes

- Un modelo DFR tiene **3 componentes**:

$$RSV_{d_j q} = \sum_{t_i \in Q} w_{ij} \quad \text{con} \quad w_{ij} = tf_{iq} \times Inf_1(tfn_{ij}) \times P_{risk}(tfn_{ij})$$

  - Inf_1, ***contenido informativo*** del término t_i en doc. d_j
  - P_{risk}, ***riesgo asumido*** al aceptar t_i como descriptor válido del doc. d_j
  - tfn_{ij}, frecuencia tf_{ij} del término t_i en doc. d_j tras ser ***normalizada respecto a longitud del doc.***

Comp. 1: Modelo Aleatorio

- **Modelo de distribución de los términos**
- $Prob_1(tf_{ij})$: probabilidad término t_i aparezca tf_{ij} veces en doc. d_j
- Inf_1, ***contenido informativo*** del término t_i en doc. d_j

$$Inf_1 = -\log_2 Prob_1$$

  - término con alta probabilidad de aparecer en un doc. (*"de no-especialidad"*): escaso contenido informativo
  - término con poca probabilidad de aparecer en un doc. (*"de especialidad"*): alto contenido informativo

Comp. 1: Ejemplos

- **Distribución binomial**:

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

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

- **Distribución geométrica**:

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

Comp. 2: Primera Normalización

- Sea un término poco común (*"de especialidad"*) que aparece en un doc. ...
  - ... muy pocas veces: puede ser por casualidad, no conviene usarlo (*riesgo alto*)
  - ... muchas veces: seguro relacionado con el tema, debemos usarlo (*riesgo bajo*)
- **Ponderar *contenido informativo* (Inf_1) respecto riesgo** al tomarlo como descriptor (P_{risk})

Comp. 2: Ejemplos

- ***Normalización L***:

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

- ***Normalización B***:

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

donde TF_i es la frecuencia total del término t_i en la colección
df_i es nº docs. que contienen el término t_i.

Comp. 3: Segunda Normalización

- **Normalizar la frecuencia** tf_{ij} del término t_i en el documento d_j en base a:
  - Longitud del documento (dl_j)
  - Longitud media de los documentos (dl_{avg})

- **Ejemplos**:

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

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

Índice

- Introducción
- Conceptos de Teoría de Probabilidades
- Principio de Ordenación por Probabilidad
- Modelo de Independencia Binaria
- Okapi BM25
- Paradigma DFR
- **Conclusión**

Conclusión

- Base formal: *teoría de probabilidades*
- *Ppo. de Ordenación por Probabilidad*
  - Ordenación por probabilidad de relevancia
  - Recuperación óptima

Conclusión (cont.)

- *Modelo de Independencia Binaria*
  - Modelo básico
- *Okapi BM25*
  - Evolución: frecuencia del término + longitud
- *Paradigma DFR*
  - *Metodología* vs. modelo
  - *Ganancia de información* vs. probabilidad de relevancia