Fundamentos de la Minería de Datos Web

Máster Online en Ciencia de Datos

UCO ONLINE

Dr. José Raúl Romero

Profesor Titular de la Universidad de Córdoba y Doctor en Ingeniería Informática por la Universidad de Málaga. Sus líneas actuales de trabajo se centran en la democratización de la ciencia de datos (*Automated ML* y *Explainable Artificial Intelligence*), aprendizaje automático evolutivo y analítica de software (aplicación de aprendizaje y optimización a la mejora del proceso de desarrollo de software).

Miembro del Consejo de Administración de la *European Association for Data Science*, e investigador senior del Instituto de Investigación Andaluz de *Data Science and Computational Intelligence*.

Director del **Máster Online en Ciencia de Datos** de la Universidad de Córdoba.

Modelos de Recuperación de la Información

UCO ONLINE

¿Qué es un modelo?

- Un **modelo de IR** predice y explica lo que un usuario encuentra relevante de una consulta (*user query*)
- La corrección de las predicciones de un modelo se prueban inicialmente en experimentos controlados
- Un modelo viene **definido por**
  - La manera en que se representan las consultas

    Subsistema de consulta
  - La manera en que se representan los documentos

    Mecanismo de indexación
  - La manera en que se realiza el emparejamiento de consultas y documentos

    Mecanismo de evaluación

¿Qué es un modelo?

- Existen diferentes **modelos de IR**
  - Modelos de emparejamiento exacto (*exact match models*)
    - Booleano
    - Modelos de regiones o booleano extendido
  - Modelos vectoriales
  - Modelos probabilísticos
  - Modelos difusos
    - Básico
    - Básico booleano
    - Consultas ponderadas por un único peso
    - Consultas ponderadas por múltiples pesos

    ...

- Cada modelo representa los documentos y consultas de forma distinta
- Todos usan el mismo marco teórico:
  - Se tratan documentos y consultas como **bolsas de palabras** ("bag" of words) o términos
  - Se ignora la posición de la sentencia en el documento

Dado un conjunto de documentos **D**, el vocabulario $\boldsymbol{V} = \{t_1, t_2, .., t_{|V|}\}$ es el conjunto distintivo de términos en **V**. Un peso $\mathbf{w_{ij}} > 0$ se asocia a cada término $\mathbf{t_i}$ de un documento $\mathbf{d_j}$ en **D** ($\mathbf{w_{ij}} = 0$ si el término $\mathbf{t_i}$ no aparece en $\mathbf{d_j}$)

$$\mathbf{d}_j = (w_{1j}, w_{2j}, \ldots, w_{|V|j})$$

Modelos de Recuperación de la Información

Modelo de IR de emparejamiento exacto > Booleano

UCO ONLINE

Modelo Booleano > Características

- Modelo más primitivo y bastante deficiente para IR
- Modelo muy popular basado en el Álgebra de Boole (marco teórico sólido y formalizado)
  - Útil para combinar con otros modelos (p.ej. para excluir documentos – Google con "AND")
- **Relevancia binaria** – un documento es relevante o no (*exact match*)
  - Consultas **AND**

    Los documentos deben tener todas las palabras
  - Consultas **OR**

    Los documentos deben contener alguna palabra
  - Consultas de **una palabra**

    El documento es relevante sí y solo sí contiene la palabra

Modelo Booleano > Base de datos

- Los documentos se representan por medio de **palabras clave** (***keywords***)
- La indexación asocia un peso binario a cada término índice
  - CERO si el término no aparece en el documento
  - UNO si el término aparece al menos una vez

$$w_{ij} = \begin{cases} 1 & \text{if } t_i \text{ appears in } \mathbf{d}_j \\ 0 & \text{otherwise.} \end{cases}$$

1	0	1	1	0	0	0

Modelo Booleano > Subsistema de consulta

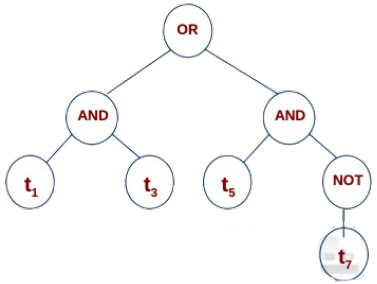
- Combinaciones de términos conectados mediante **operadores lógicos**

$$(t_1 \text{ OR } t_2) \text{ AND } (t_1 \text{ AND NOT } t_5)$$

- No permite ponderación de términos de consulta
- Se suelen normalizar las consultas para que resulten más eficientes (FNC y FND)

Modelo Booleano > Mecanismo de evaluación

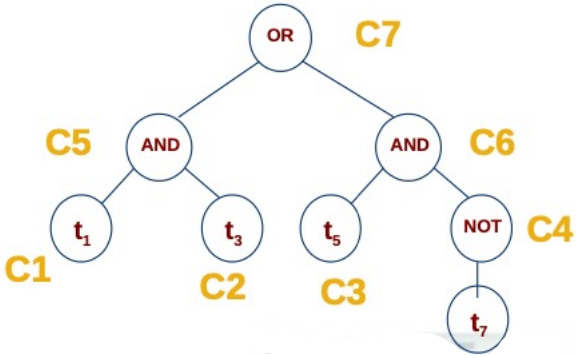
- El **grado de similitud** entre un documento y una consulta será binario
  - Un documento es relevante si su grado de similitud es UNO
  - En otro caso, es irrelevante
- Representación de **consultas como árboles binarios**
  - Las hojas son términos índice
  - Los nodos intermedios son operadores



- **Procedimiento de abajo a arriba**. Primero se evalúan los átomos y luego las expresiones, aplicando la acción de los operadores:
  - AND → Intersección
  - OR → Unión

Modelo Booleano > Mecanismo de evaluación

- Ejemplo:
  - **C1** – Documentos que contengan el término t_1
  - **C2** – Documentos que contengan el término t_3
  - **C3** – Documentos que contengan el término t_5
  - **C4** – Documentos que contengan el término t_7
  - **C5** – Intersección **C1** y **C2**
  - **C6** – Intersección **C3** y **C4**
  - **C7** – Unión **C5** y **C6**



RSV = 0 (ausencia)
RSV = 1 (presencia)

Emparejamiento exacto

Modelo Booleano > Inconvenientes

- Criterio de recuperación muy tajante
  - Un documento es relevante o no
  - No hay grados de relevancia
- Operadores demasiado rígidos
  - Demasiado restrictivo (AND)
  - Demasiado inclusivo (OR)
- No permite ordenar la salida
- No suele utilizar *feedback* para mejorar la salida
- Da lo mismo que un documento contenga 1 o 100 veces las *keywords* de consulta

Modelos de Recuperación de la Información

Modelo de IR de emparejamiento exacto > Vectorial

UCO ONLINE

Modelo Vectorial > Características

- Modelo propuesto a finales de los 60 pero posiblemente el más utilizado
- Cada documento es un vector de pesos en el dominio de los reales que podrá situarse en un espacio vectorial *n*-dimensional
  - Los documentos están bien definidos
  - Se crean grupos de documentos próximos entre sí

    Clúster de documentos

    Relevantes para la misma necesidad de información
  - Cálculo muy rápido del RSV

    Los documentos están agrupados por su grado de semejanza

Modelo Vectorial > Base de datos

- Documentos = Vectores de n-valores reales
  - Cada valor indica la importancia relativa del término en el documento
  - Cada valor puede estar normalizado o no

0.5	0	0.7	0.9	0.1	0	0.2

- Mecanismo de indexación basado en alguna variación de TF o TF-IDF (peso $\mathbf{w_{ij}}$ de un término $\mathbf{t_i}$ del documento $\mathbf{d_j}$)

- Esquema de frecuencia de términos (TF) – $\mathbf{w_{ij}}$ es el número de veces que $\mathbf{t_i}$ aparece en $\mathbf{d_j}$, denotado por $\mathbf{f_{ij}}$ (*se puede normalizar*)

Modelo Vectorial > Base de datos

- **Esquema TF-IDF** (frecuencia de términos – frecuencia de documento inversa) – hay multitud de variantes de esta representación, pero veremos la más sencilla

  N = Número total de documentos

  df_i = Número de documentos en los que aparece **t_i**

  f_{ij} = Frecuencia absoluta de **t_i** en **d_j**

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \ldots, f_{|V|j}\}}$$

Frecuencia de términos normalizada
Si **t_i** no está en el documento, entonces **tf_{ij}** = 0

$$idf_i = \log \frac{N}{df_i}$$

Frecuencia del documento inversa del término **t_i**

Si un término aparece en un número elevado de documentos, probablemente no será importante ni discriminatorio

$$w_{ij} = tf_{ij} \times idf_i$$

Peso del término TD-IDF
(es habitual normalizar este valor)

Modelo Vectorial > Subsistema de consulta

- La consulta se representa exactamente igual que el documento

  Consulta = vector n-dimensional de valores reales

- La mayoría de los pesos serán 0 – solo se indican aquellos con peso distinto a cero
- Los términos no están conectados por ningún operador – la consulta es un todo

1	0	0.8	0	0.2	0	0

- Salton & Buckley sugieren la siguiente formulación para el peso del término $\mathbf{t_i}$ en una consulta **q**:

$$w_{iq} = \left(0.5 + \frac{0.5 f_{iq}}{\max\{f_{1q}, f_{2q}, \ldots, f_{|V|q}\}}\right) \times \log \frac{N}{df_i}$$

TF IDF

Modelo Vectorial > Mecanismo de evaluación

- El **grado de similitud** entre un documento (**d**) y una consulta (**q**)

  Se toman sus representaciones vectoriales:

  $d = (d_1, d_2, \ldots, d_n)$

  $q = (q_1, q_2, \ldots, q_n)$

- El valor RSV será mayor cuanto más similares sean **d** y **q**

Producto escalar

$$RSV(d,q) = \sum_{i=1}^{n} d_i * q_i$$

Distancia euclídea

$$RSV(d,q) = -\sqrt{\sum_{i=1}^{n} |d_i - q_i|^2}$$

Distancia Manhattan

$$RSV(d,q) = -\sum_{i=1}^{n} |d_i - q_i|$$

Medidas métricas (proximidad en el espacio documental)

Coseno

$$RSV(d,q) = \frac{\sum_{i=1}^{n} d_i * q_i}{\sqrt{\sum_{i=1}^{n} d_i^2} * \sqrt{\sum_{i=1}^{n} q_i^2}}$$

Dice

$$RSV(d,q) = \frac{2 * \sum_{i=1}^{n} d_i * q_i}{\sum_{i=1}^{n} (d_i^2 + q_i^2)}$$

Jaccard

$$RSV(d,q) = \frac{\sum_{i=1}^{n} d_i * q_i}{\sum_{i=1}^{n} (d_i^2 + q_i^2 - d_i * c_i)}$$

Medidas angulares (misma dirección)

Modelo Vectorial > Pros y Cons

Ventajas del modelo

- Permite hacer correspondencias parciales
- Ordena los resultados por similitud
- El tamaño de salida controlable por el usuario:
  - Poniendo límite al número de documentos recuperados
  - Aceptando aquellos documentos que superan un umbral

Inconvenientes del modelo

- No incorpora la noción de correlación entre términos

Modelos de Recuperación de la Información

Modelo de IR de emparejamiento exacto > Probabilístico

Fuente: http://ccdoc-tecnicasrecuperacioninformacion.blogspot.com/2012/12

Modelo Probabilístico > Características

- Modelo propuesto a finales de los 70, conocido como modelo de independencia binaria (BIR)
- Cada documento es un vector binario para presencia (1) o ausencia (0) de términos

1	0	1	1	0	0	0

CDR = Conjunto de documentos relevantes

CDI = Conjunto de documentos irrelevantes

- Los documentos se clasifican en relevantes o irrelevantes
- Existe una respuesta ideal del sistema

  Conjunto de documentos relevantes (= Conjunto de respuesta ideal)
- Existe una consulta ideal

  Proporciona un conjunto de respuesta ideal

  Se desconoce a priori qué términos deberían aparecer

Premisas de funcionamiento del modelo

Modelo Probabilístico > Objetivo

Tomar una consulta del usuario y refinarla hasta el conjunto de respuesta ideal

- Reformulación sucesiva de los términos de la consulta mediante su ponderación
- Presencia (1) * Peso (w)
- Acercar la consulta inicial a la consulta ideal

Modelo Probabilístico > Ponderación de términos

- El **proceso de ponderación** de los términos de la consulta es el cálculo de probabilidad de que exista dicho término en el CDR y la probabilidad de que se encuentre presente en el CDI

$P(Ti / R)$	Probabilidad de que un término se encuentre en el **Conjunto de Documentos relevantes CDR**	0,95	Término muy relevante
$P(Ti / R\neg)$	Probabilidad de que un término se encuentre en el **Conjunto de Documentos Irrelevantes CDI**	0,02	Término poco relevante

El cálculo del peso para el término t_i se calcula con la **razón de Odds** aplicado a la consulta del usuario

$$W_{(Ti)} = \frac{P(Ti / R)}{P(Ti / R\neg)}$$

- **W = Weight = Peso de un término Ti**
- P(Ti/R) = Probabilidad de que el término esté presente en el **conjunto de documentos relevantes CDR**
- P(Ti/R¬) = Probabilidad de que el término esté presente en el **conjunto de documentos irrelevantes CDI**

Modelo Probabilístico > Ponderación de términos

- Inicialmente se desconoce el número de documentos que forman el CDR y el CDI
- Se otorga unos valores iniciales, llamados de "**Máxima incertidumbre**"

Valor de máxima incertidumbre

$$P(Ti / R) = 0{,}5$$

Nº de documentos con el término Ti

Nº total de documentos de la colección

$$P(Ti / R\neg) = ni / N$$

- Existen distintos factores que pueden afectar al peso final $\mathbf{w_{ti}}$ del término:
  - Presencia o ausencia de los términos en la consulta
  - Independencia de la distribución de los términos en el CDR

Método estándar de cálculo de pesos de términos de la consulta en modelo probabilítico de independencia binaria

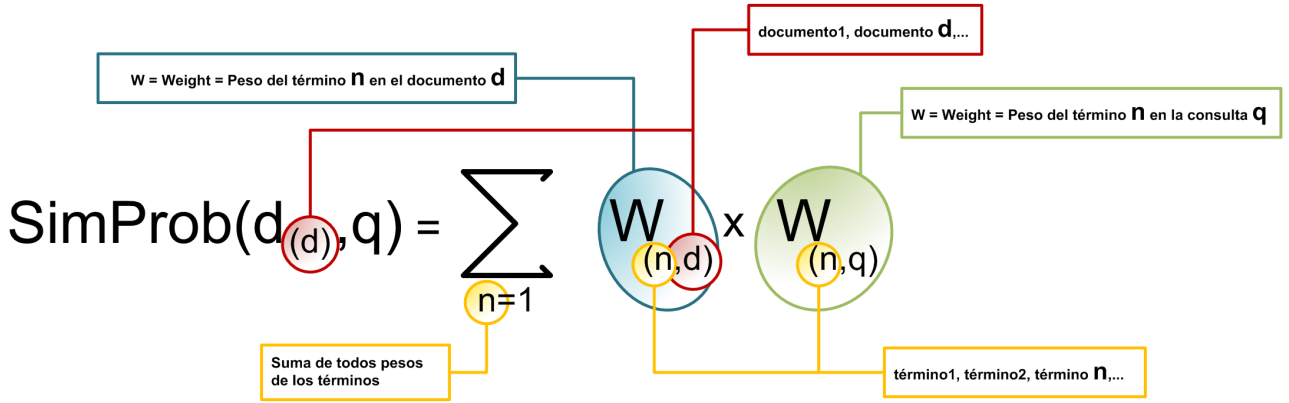
$$W_{(Ti)} = \log_{10} \frac{P(Ti / R)}{1 - P(Ti / R)} + \log_{10} \frac{1 - P(Ti / R\neg)}{P(Ti / R\neg)}$$

Prob. de presencia y ausencia en CDR

Prob. de presencia y ausencia en CDI

Modelo Probabilístico > Mecanismo de evaluación

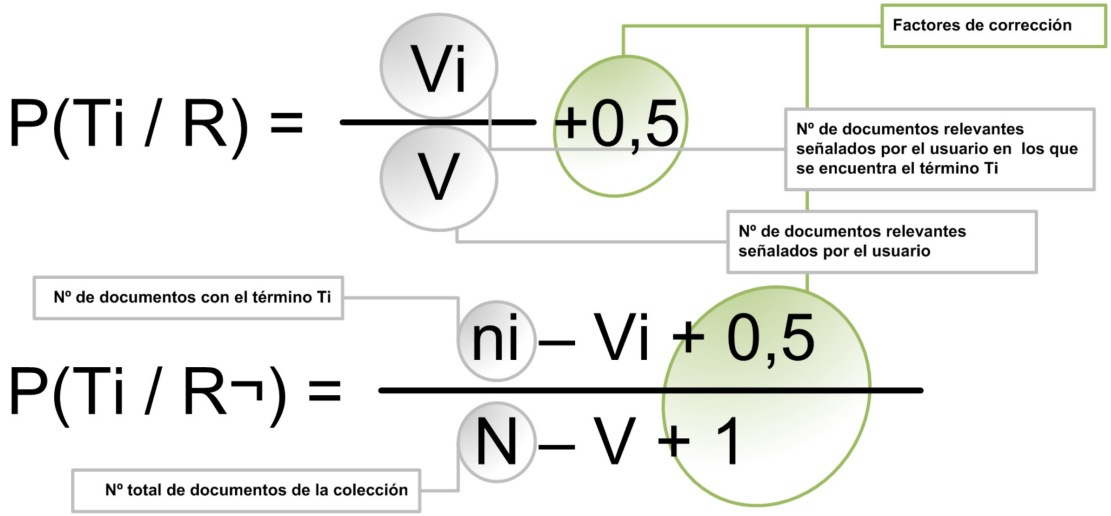
- El grado de similitud entre un documento (**d**) y una consulta (**q**) basado en producto escalar



- El sistema ordena los documentos de la colección conforme al orden decreciente de su probabilidad de relevancia (**`SimProb`**) con respecto a la consulta del usuario
  - Se mostrará en primer lugar el documento con **`SimProb`** más alta

Modelo Probabilístico > Mecanismo de evaluación

- Una vez mostrados los documentos, el sistema pide al usuario que señale la relevancia de los documentos → reajuste de CDR y CDI
- Se reformulan las probabilidades de que un término esté tanto en el CDR como en el CDI (”retroalimentación de relevancia”) – el proceso se suele repetir 1-2 ciclos



Modelo Probabilístico > Pros y Cons

Ventajas del modelo

- Retroalimentación por relevancia (*feedback*)
- Asume independencia de términos de la consulta
- Considerado uno de los mejores modelos dados sus resultados con colecciones reales y corpus de entrenamiento
- Método de recuperación mediante equiparación parcial, superando a la equiparación exacta del modelo booleano

Inconvenientes del modelo

- Mantiene el modelo binario de recuperación de información
- Demanda computacional
- Necesita efectuar hipótesis inicial, que puede no ser acertada
- No tiene en cuenta la frecuencia de aparición de cada término en el documento (frente vectorial)

¡Gracias!

UCO ONLINE