Modelos de Recuperación de la Información

Slide 1

Modelo booleano

- documento = conjunto de términos
- interrogación = fórmula booleana, no forma parte del espacio de documentos
- respuesta es un conjunto de documentos sin ranking

Slide 2

Modelo vectorial

- documento = vector con un peso asignado a cada término.
- interrogación = documento
- respuesta es un ranking del espacio de docuentos.

Slide 3

Modelo probabilístico

- Se presupone que existe exactamente un subconjunto R de documentos que son relevantes para una consulta dada.
- Para cada documento se intenta evaluar la probabilidad de que el usuario lo considere relevante.
- La relevancia de un documento d de calcula como

$$\frac{Pr(d \text{ relevante para } q)}{Pr(d \text{ no relevante para } q)}$$

Slide 4

- Documento= lista de términos (como en el modelo booleano)
- Interrogación (query) es un documento
- Objetivo: mediante la interacción con el usuario obtener una descripción probabilistica correcta del conjunto R.

Slide 5

Expresión en función de los términos

- $Pr(R|d_i)$ la probabilidad de que d_i sea relevante y $Pr(\overline{R}|d_j)$ la probabilidad de que d_j no sea relevante
- $Pr(R)$ la probabilidad de que un documento elegido al azar sea relevante
- $Pr(d_i|R)$ probabilidad de elegir d_i entre los documentos relevantes

Utilizando la fórmula de Bayes:

Slide 6

$$\mathsf{sim}(d_i, q) = \frac{Pr(R|d_i)}{Pr(\overline{R}|d_i)} = \frac{Pr(d_i|R) \times Pr(R)}{Pr(\overline{d_i}|\overline{R}) \times Pr(\overline{R})} \sim \frac{Pr(d_i|R)}{Pr(\overline{d_i}|\overline{R})}$$

La última simplificación se puede hacer ya que es un factor idéntico para todos los documentos y no modifica el ranking.

Sin embargo es necesario reescribir todas las probabilidades en términos de términos de índice y documentos

- $Pr(t_r|R)$ es la probabilidad de que el término t_r aparezca en un documento relevante y $Pr(\overline{t}_r|R)$ es la probabilidad de que el término t_r no aparezca en un documento relevante.
- $Pr(t_r|\overline{R})$ la probabilidad de que el término t_r aparezca en un documento no relevante.
- $w_{r,i}$ es 1 si t_r aparece en d_i, 0 en caso contrario.

Slide 7

Si asumimos independencia de aparición entre términos índice

$$Pr(d_j|R) = \left(\prod_{g_i(\vec{d}_j)=1} Pr(t_i|R)\right)\left(\prod_{g_i(\vec{d}_j)=0} Pr(\overline{t}_i|R)\right)$$

Y una fórmula similar para el denominador.

Despues de unos cuantos cálculos y tomando logaritmos se llega a

$$\mathsf{sim}(d_i, q) \sim \sum_{r=1}^{k} w_{r,i} \times w_{r,q} \times \left(\log \frac{Pr(t_r|R)}{1 - Pr(t_r|R)} + \log \frac{1 - Pr(t_r|\overline{R})}{Pr(t_r|\overline{R})}\right)$$

$$\sum_{t_r \in q \cap d_i} \left(\log \frac{Pr(t_r|R)}{1 - Pr(t_r|R)} + \log \frac{1 - Pr(t_r|\overline{R})}{Pr(t_r|\overline{R})}\right)$$

Slide 8

Como estimar las probabilidades

Se hace de forma iterativa

- Fijar una distribucion inicial.
- Despues de una iteración se recuperan V documentos relevantes; sea v_r el número de documentos en V que contienen el término t_r. Se recalcula $Pr(t_r|R)$ y $Pr(t_r|\overline{R})$ utilizando v_r.

Slide 9

Se reitera el proceso, idealmente hasta converger a una distribución "estacionaria", que deberia ser la correcta.

Alternativamente se puede interactuar con el usuario para calcular grado de relevancia de los documentos recuperados en algún(os) pasos.

Distribución inicial

Hipótesis simples:

- $Pr(t_r|R)$ es constante para todos los términos.
- $Pr(t_r|\overline{R})$ se puede aproximar por la distribución de términos en todos los documentos.

Es decir, por ejemplo,

Slide 10

$$Pr(t_r|R) = 0.5 \text{ y } Pr(t_r|\overline{R}) = n_r/N.$$

Mejora

Hemos recuperado V documentos relevantes: corresponden a los v documentos con mayor probabilidad (o preguntando al usuario). Donde v es un umbral prefijado.

Sea v el número de documentos en V y sea v_r el número de documentos en V que contienen el término t_r.

Opción 1:

Slide 11

$$Pr(t_r|R) = \frac{v_r}{v} \text{ y } Pr(t_r|\overline{R}) = \frac{n_r - v_r}{N - v}$$

Puede dar problemas si v y/o v_i son pequeños

Opción 2:

$$Pr(t_r|R) = v_r + 0.5v + 1 \text{ y } Pr(t_r|\overline{R}) = \frac{n_r - v_r + 0.5}{N - v + 1}$$

El factor de ajuste de 0.5 puede no ser muy fino

Opción 3:

Slide 12

$$Pr(t_r|R) = v_r + 0.5v + 1 \text{ y } Pr(t_r|\overline{R}) = \frac{n_r - v_r + 0.5}{N - v + 1}$$

El modelo probabilístico recupera, en teoría, los documentos con más probabilidad de ser relevantes. Sin embargo no es muy popular

- Se debe empezar adivinando y refinar iterativamente
- El modelo no tiene en cuenta la frecuencia de aparición de un término, el espacio de documentos es similar al del modelo booleano.
- Se necesita que los términos sean independientes.
- Existen estudios que muestran que es inferior al modelo vectorial.
- Tiene una base teórica distinta a la del modelo vectorial y permite extensiones que sí son populares.

Slide 13

Otros modelos

Estos modelos son, en general, bastante costosos de implementar. No siempre dan grandes resultados.

- Extensiones al modelo booleano
  - Booleano extendido
  - Conjuntos difusos
- Extensiones al modelo vectorial
  - Vectorial generalizado
  - Latent semantic indexing
  - Redes neuronales
- Extensiones al modelo probabilístico
  - Redes bayesianas
  - Redes de Inferencia Bayesianas

Slide 14

Booleano extendido

Intenta incorporar algún mecanismo de valoración al modelo booleano.

Por ejemplo si la pregunta es

crisis y misiles

los documentos que contienen uno de los dos términos parecen más relevantes que los que no contienen ninguna de las dos.

Slide 15

Supongamos que solo hay 2 términos t_x y t_y

documento d_j se corresponde con un vector de dos dimensiones $(w_{x,j}, w_{y,j})$

con pesos normalizados para que esten entre 0 y 1.

Por ejemplo $w_{x,j} = f_{x,j}\frac{idf_x}{\max_r idf_r}$ donde

$f_{x,j}$ es la frecuencia normalizada del término t_x en d_j

idf_x el inverso de la frecuencia de docuemntos con el término t_x.

Slide 16

$d = (x, y)$

para $q_{or} = t_x \vee t_y$ el documento más deseable es (1,1), medimos que esté alejado de (0,0)

para $q_{and} = t_x \wedge t_y$ el documento más relevante es (1,1), medimos que esté proximo a (1,1)

Slide 17

$$\mathsf{sim}(q_{or}, d) = \sqrt{\frac{x^2 + y^2}{2}}$$

$$\mathsf{sim}(q_{and}, d) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}}$$

Idea es extender a k dimensiones y utilizar las mismas funciones de similaridad (utilizando una norma en un espacio multidimensional.

Aplicación adicional es el tratamiento de fórmulas más complicada

$$q = ((t_1 \wedge t_2) \vee (t_3 \wedge t_4)) \wedge t_5$$

Se calcula el peso asociado a cada pregunta simple y se itera aplicando las fórmulas.

Slide 18

Conjuntos difusos

La lógica difusa intenta representar clases cuyas fronteras no están bien definidas.

A cada conjunto A se le asocia una función de pertenencia $\mu_A : U \rightarrow [0,1]$.

Se extiende esta función con operación mediante operaciones booleanas

Slide 19

$$\begin{aligned}\mu_{\overline{A}}(u) &= 1 - \mu_A(u)\\ \mu_{A\cup B}(u) &= \max\{\mu_A(u), \mu_B(u)\}\\ \mu_{A\cap B}(u) &= \min\{\mu_A(u), \mu_B(u)\}\end{aligned}$$

Utilizar un tesaurus que define relación entre términos

Expandir la pregunta con términos relacionados de manera que otros documentos relevantes puedan ser recuperados.

Un tesauros se puede definir midiendo la correlación entre pares de términos: $c_{i,j}$ es la correlación entre los término t_i y t_j

$c_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{ij}}$

$n_{i,j}$ es el número de documentos que contiene t_i y t_j.

Slide 20

Con esta valoración la función de pertenencia de un documento se define como

$$\mu_j t_r = 1 - \prod_{t_l \in d_j} (1 - c_{r,l}$$

Así se asocia un conjunto difuso con cada término, el de aquellos documentos cuya función de pertenencia sobrepasa un determinado umbral.

Se compone el conjunto final como en el modelo booleano.

Slide 21

Vectorial generalizado

- Asociar un vector o un conjunto de vectores con los términos índice
- Por ejemplo los 2^t vectores que representan todas las combinaciones de ocurrencias de términos en un documento
- Utilizar este conjunto como base de un espacio vectorial de términos
- El modelo tiene en cuenta la co-ocurrencia de términos

Slide 22

- Un documento se representa mediante un vector (combinación lineal de los vectores asociados a los términos)
- Una interrogación es un documento
- La medida de similaridad se define como en el modelo vectorial

Slide 23

Latent semantic indexing

- Idea: mapear el espacio de vectores a uno de menor dimensión, conservando lo mejor posible las distancias entre vectores
- Se colapsan vectores que son relativamente dependientes en una coordenada
- Los vectores en el espacio reducido representan conceptos en lugar de términos
- Al preguntar por un término se recuperan documentos relevantes al concepto asociado, aún cuando el término no aparezca en el documento.

Slide 24

Redes neuronales

Slide 25

- La idea general es tratar de expandir los términos de la consulta por iteración
- Generamos una red neuronal en la que los nodos son los términos de la interrogación, los términos de los documentos y los documentos en sí. Las flechas representan una potencial activación.
- Los términos de la interrogación activan los mismos términos de los documentos.
- Los términos de los documentos activan los documentos donde aparecen.
- Los documentos activan todos los términos que contienen.
- El proceso itera la activación, con intensidades decrecientes hasta converger.
- El nivel alcanzado por cada documento es su relevancia.
- El proceso permite que se activen documentos en forma indirecta, cuándo contienen un término que en los documentos de la colección aparece frecuentemente junto a otro de la interrogación.