Recuperació de la Informació (REIN)

Grau en Enginyeria Informàtica

Departament de Ciències de la Computació (CS)

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Escola Politècnica Superior d'Enginyeria
de Vilanova i la Geltrú

Índex

Models de Recuperació de la Informació, I

Què és un model de Recuperació de la Informació?

Determina:

- El conjunt de vistes lògiques (o representacions) dels documents
  (quina info s'emmagatzema/indexa de cada document?),
- un llenguatge d'interrogació
  (quin tipus de consultes podran fer-se?),
- un criteri de rellevància
  (donada una consulta, què es fa amb cada document?).

Models de Recuperació de la Informació, I

Dos models de RI, entre d'altres

En aquesta assignatura:

- Model booleà
  - El més senzill, poc usat.
  - Les consultes són expressions booleanes, les respostes són exactes.
  - Extensió: consultes per frase.
- Model vectorial
  - Pesos en els termes i en els documents.
  - Grau de similitud entre consultes i documents, respostes aproximades, ordenació.

Índex

Model booleà de RI

Assumeix rellevància binària

Documents:

En aquest model, cada document s'identifica pel conjunt de termes que conté.

- L'ordre en què apareixen és irrellevant.
- El nombre de vegades que es repeteixen és irrellevant (però hi ha un model molt similar, anomenat bag-of-words o BoW, que considera rellevants les freqüències).

Model booleà de RI

Assumeix rellevància binària

Documents:

En aquest model, cada document s'identifica pel conjunt de termes que conté.

- L'ordre en què apareixen és irrellevant.
- El nombre de vegades que es repeteixen és irrellevant (però hi ha un model molt similar, anomenat bag-of-words o BoW, que considera rellevants les freqüències).

Per tant, per un conjunt de termes $\mathcal{T} = \{t_1, \ldots, t_T\}$, un document és un subconjunt de $\mathcal{T}$.

Cada document és representat com un vector de bits (0 o 1) de longitud T, $d = (d_1, \ldots, d_T)$, on

- $d_i = 1$ si i només si t_i apareix en d,
- $d_i = 0$ si i només si t_i no apareix en d.

Consultes en el model booleà, I

Consultes booleanes, respostes exactes

Consultes atòmiques:

D'una sola paraula. La resposta és el conjunt de documents que la contenen.

Consultes compostes:

S'usen els operadors de l'àlgebra de Boole, excepte la negació:

- OR, AND: funcionen com la unió o intersecció de respostes.
- El conjunt diferència, resta o "complement", s'usa en lloc de la negació unària:

  t_1 BUTNOT t_2, correspon a t_1 AND NOT t_2;
- Motivació: evitar generar conjunts de resposta amb una mida intractable.

Exemple, I

Considerem els 7 documents següents amb un vocabulari de 6 termes.

d1 = un tres

d2 = dos dos tres

d3 = un tres quatre cinc cinc cinc

d4 = un dos dos dos dos tres sis sis

d5 = tres quatre quatre quatre sis

d6 = tres tres tres sis sis

d7 = quatre cinc

Exemple, II

Els documents en el model booleà

$$
\begin{array}{lc}
 & \begin{array}{cccccc} cinc & dos & quatre & sis & tres & un \end{array} \\
d1 = & \left[\begin{array}{cccccc} 0 & 0 & 0 & 0 & 1 & 1 \end{array}\right] \\
d2 = & \left[\begin{array}{cccccc} 0 & 1 & 0 & 0 & 1 & 0 \end{array}\right] \\
d3 = & \left[\begin{array}{cccccc} 1 & 0 & 1 & 0 & 1 & 1 \end{array}\right] \\
d4 = & \left[\begin{array}{cccccc} 0 & 1 & 0 & 1 & 1 & 1 \end{array}\right] \\
d5 = & \left[\begin{array}{cccccc} 0 & 0 & 1 & 1 & 1 & 0 \end{array}\right] \\
d6 = & \left[\begin{array}{cccccc} 0 & 0 & 0 & 1 & 1 & 0 \end{array}\right] \\
d7 = & \left[\begin{array}{cccccc} 1 & 0 & 1 & 0 & 0 & 0 \end{array}\right]
\end{array}
$$

(Inventeu algunes consultes i calculeu les seves respostes.)

Consultes en el model booleà, II

Els resultats no estan ordenats

Els resultats no estan quantificats:

Un document o bé

- coincideix amb la consulta
  (és totalment rellevant),
- o bé no coincideix amb la consulta
  (és totalment irrellevant).

Consultes en el model booleà, II

Els resultats no estan ordenats

Els resultats no estan quantificats:

Un document o bé

- coincideix amb la consulta
  (és totalment rellevant),
- o bé no coincideix amb la consulta
  (és totalment irrellevant).

Depenent de les necessitats de l'usuari i de l'aplicació, això pot ser bo o dolent.

- Busco un document més o menys conegut?
- Busco informació relacionada?

Consultes per frase, I

Estenent el model booleà

Consultes per frase: conjunció més adjacència

Respondre amb el conjunt de documents que presenten els termes de la consulta de forma consecutiva (apareixen junts).

- Si l'usuari consulta "George Harrison" no vol un document que mencioni tant en George Orwell com en Richard Harrison.
- El model booleà no permet consultes per frase.

Consultes per frase, II

Possibles solucions

Opcions:

- Executar una consulta conjuntiva, després tornar a repassar el conjunt de resposta per eliminar els casos de no adjacència.

  Aquesta opció pot ser molt lenta en casos on hi hagi molts “falsos positius”.

- Emmagatzemar dins l’índex informació addicional sobre l’adjacència de qualsevol parell de termes d’un document (p.e. posicions).
- Emmagatzemar dins l’índex informació addicional sobre la possibilitat de “parelles interessants” de paraules.

Índex

El model vectorial de RI, I

Base de totes les aproximacions exitoses

- El nombre d'ocurrències d'un terme en un document és rellevant.
- Segueix sent irrellevant l'ordre en què apareixen els termes en el document.
- No tots els termes són igual d'importants.
- Per un conjunt de termes $\mathcal{T} = \{t_1, \ldots, t_T\}$, un document és $d = (w_1, \ldots, w_T)$ de reals en lloc de bits.
- w_i és el pes de t_i en el document d.

El model vectorial de RI, II

Els termes formen T dimensions, els documents $\in IR^T$

- Un document és un vector $\in IR^T$.
- La col·lecció de documents es converteix en una matriu:

  termes $\times$ documents

  però mai calculem explícitament aquesta matriu.
- Les consultes també són representades com a vectors $\in IR^T$.

Esquema de pesos *tf-idf*, I

Com assignar un vector de pesos a un document

Dos principis:

- Com més freqüent és un terme t en un document d, més pes hauria de tenir.

Esquema de pesos *tf-idf*, I

Com assignar un vector de pesos a un document

Dos principis:

- Com més freqüent és un terme t en un document d, més pes hauria de tenir.
- Però si un terme és molt freqüent en tots els documents de la col·lecció, serveix menys per discriminar-los i llavors voldríem decrementar el seu pes.

Esquema de pesos *tf-idf*, II

Un document és un vector de pesos:

$$d = [w_{d,1}, \ldots, w_{d,i}, \ldots, w_{d,T}]$$

Cada pes és el producte de dos valors:

$$w_{d,i} = tf_{d,i} \cdot idf_i$$

La freqüència dels termes tf és:

$tf_{d,i} = \frac{f_{d,i}}{\max_j f_{d,j}}$, on $f_{d,j}$ és la freqüència de t_j en d.

I la freqüència inversa dels documents idf és:

$$idf_i = \log_2 \frac{D}{df_i}$$

on $D =$ nombre de documents i $df_i =$ nombre de documents que contenen el terme t_i.

Exemple, I

	$cinc$	dos	$quatre$	sis	$tres$	un	**maxf**
$d1 =$	0	0	0	0	1	1	1
$d2 =$	0	2	0	0	1	0	2
$d3 =$	3	0	1	0	1	1	3
$d4 =$	0	4	0	2	1	1	4
$d5 =$	0	0	3	1	1	0	3
$d6 =$	0	0	0	2	3	0	3
$d7 =$	1	0	1	0	0	0	1
df $=$	2	2	3	3	6	3	

Exemple, II

$$
\begin{array}{lcccccccc}
\textbf{df} = & & 2 & 2 & 3 & 3 & 6 & 3 & \\
d3 = & [& 3 & 0 & 1 & 0 & 1 & 1 &] \\
d3 = & [& \frac{3}{3}\log_2\frac{7}{2} & \frac{0}{3}\log_2\frac{7}{2} & \frac{1}{3}\log_2\frac{7}{3} & \frac{0}{3}\log_2\frac{7}{3} & \frac{1}{3}\log_2\frac{7}{6} & \frac{1}{3}\log_2\frac{7}{3} &] \\
= & [& 1.81 & 0 & 0.41 & 0 & 0.07 & 0.41 &]
\end{array}
$$

$$
\begin{array}{lcccccccc}
d4 = & [& 0 & 4 & 0 & 2 & 1 & 1 &] \\
d4 = & [& \frac{0}{4}\log_2\frac{7}{2} & \frac{4}{4}\log_2\frac{7}{2} & \frac{0}{4}\log_2\frac{7}{3} & \frac{2}{4}\log_2\frac{7}{3} & \frac{1}{4}\log_2\frac{7}{6} & \frac{1}{4}\log_2\frac{7}{3} &] \\
= & [& 0 & 1.81 & 0 & 0.61 & 0.06 & 0.31 &]
\end{array}
$$

Similitud de documents en el model vectorial, I

La mesura de similitud cosinus

- Pot passar que “vectors similars” tinguin longituds diferents.

Similitud de documents en el model vectorial, I

La mesura de similitud cosinus

- Pot passar que “vectors similars” tinguin longituds diferents.
- Millor si comparem només les seves direccions.

Similitud de documents en el model vectorial, I

La mesura de similitud cosinus

- Pot passar que "vectors similars" tinguin longituds diferents.
- Millor si comparem només les seves direccions.
- O, el que és el mateix, els normalitzem abans de comparar-los perquè tinguin la mateixa longitud euclidiana.

$$sim(d1, d2) = \frac{d1 \cdot d2}{|d1||d2|} = \frac{d1}{|d1|} \cdot \frac{d2}{|d2|}$$

on

$$v \cdot w = \sum_i v_i \cdot w_i, \text{ and } |v| = \sqrt{v \cdot v} = \sqrt{\sum_i {v_i}^2}$$

Similitud de documents en el model vectorial, I

La mesura de similitud cosinus

- Pot passar que "vectors similars" tinguin longituds diferents.
- Millor si comparem només les seves direccions.
- O, el que és el mateix, els normalitzem abans de comparar-los perquè tinguin la mateixa longitud euclidiana.

$$sim(d1, d2) = \frac{d1 \cdot d2}{|d1||d2|} = \frac{d1}{|d1|} \cdot \frac{d2}{|d2|}$$

on

$$v \cdot w = \sum_i v_i \cdot w_i, \text{ and } |v| = \sqrt{v \cdot v} = \sqrt{\sum_i {v_i}^2}$$

- Els pesos són tots no negatius.
- Per tant, tots els cosinus / similituds estan entre 0 i 1.

Similitud de documents en el model vectorial, II

Exemple: similitud del cosinus

$$d3 = [\ 1.81 \quad 0 \quad 0.41 \quad 0 \quad 0.07 \quad 0.41\ ]$$
$$d4 = [\ 0 \quad 1.81 \quad 0 \quad 0.61 \quad 0.06 \quad 0.31\ ]$$

Aleshores

$$|d3| = 1.898, \quad |d4| = 1.933, \quad d3 \cdot d4 = 0.129$$

i $sim(d3, d4) = 0.035$, o sigui, que tenen una similitud baixa.

Resposta a les consultes

- Les consultes també poden transformar-se en vectors.

Resposta a les consultes

- Les consultes també poden transformar-se en vectors.
- De vegades, s'usen pesos *tf-idf*; molt sovint s'usen pesos binaris.

Resposta a les consultes

- Les consultes també poden transformar-se en vectors.
- De vegades, s'usen pesos *tf-idf*; molt sovint s'usen pesos binaris.
- $sim(doc, consulta) \in [0, 1]$.

Resposta a les consultes

- Les consultes també poden transformar-se en vectors.
- De vegades, s'usen pesos *tf-idf*; molt sovint s'usen pesos binaris.
- $sim(doc, consulta) \in [0, 1]$.
- Resposta: llista de documents en ordre decreixent de similitud.

Resposta a les consultes

- Les consultes també poden transformar-se en vectors.
- De vegades, s'usen pesos *tf-idf*; molt sovint s'usen pesos binaris.
- $sim(doc, consulta) \in [0, 1]$.
- Resposta: llista de documents en ordre decreixent de similitud.
- Veurem que també ens pot ser útil per comparar $sim(d1, d2)$.