INFORMATION RETRIEVAL

venerdì 3 novembre 2023 17:15

Definizione
Il recupero delle informazioni (IR) consiste nel trovare materiale (solitamente documenti) di un
natura non strutturata (solitamente testo) che soddisfa un bisogno informativo
da grandi raccolte (solitamente archiviate su computer)
Centinaia di milioni di persone si impegnano ogni giorno nel recupero delle informazioni giorno in cui utilizzano un motore di ricerca Web o eseguono ricerche nella posta elettronica

Il sistema IR aiuta gli utenti a trovare le informazioni di cui hanno bisogno
ma non restituisce esplicitamente le risposte alla domanda. Avvisa
riguardanti l'esistenza e l'ubicazione dei documenti che potrebbero consistere
delle informazioni richieste.
Un sistema IR ha la capacità di rappresentare, archiviare, organizzare e accedere elementi informativi. Per effettuare la ricerca è necessario un insieme di parole chiave

Differenze:
Information retrvieval:
-Il programma software che si occupa con l'organizzazione, lo stoccaggio, recupero e valutazione di informazioni dal documento repository particolarmente testuali informazione.
-Recupera informazioni su delle materie.
-È probabile che piccoli errori scompaiano inosservato.
-Non strutturato
-Ambiguo
-I risultati ottenuti sono corrispondenze approssimative
-I risultati sono ordinati per pertinenza

Data retrieval:
-È un processo di identificazione e recuperare i dati da database, in base alla query forniti dall'utente o dall'applicazione.
-Determina le parole chiave in query dell'utente e recupera i dati.
-Un singolo oggetto in errore significa totale fallimento.
-Strutturato
-Semantica ben definita

-I risultati ottenuti sono esatti
-I risultati non sono ordinati per
pertinenza

Formalmente..
Vocabolario $V = \{w1, w2, \dots, wN\}$
Query $Q = q1, q2, \dots, qm\ where\ qi \in V$
Documento $dk = \{dk1, dk2, \dots, dkm\}\ where\ dki \in V$
Collezione C = $\{d1, d2, \dots, dn\}$
Insieme di documenti rilevanti R q $\subseteq C$

Il modello booleano
Il modello booleano è un modello deterministico che utilizza operatori logici per combinare i termini di una query e restituire i documenti che soddisfano la condizione logica.
Operatori usati:
AND
OR
NOT

Limitazioni
Sono molto rigidi e difficili da esprimere richieste complesse degli utenti.
Sono difficili da controllare il numero e la qualità dei documenti recuperati.
Sono difficili da fornire feedback sulla pertinenza.

Il modello dello spazio vettoriale
La rappresentazione di un insieme di documenti come vettori in uno spazio vettoriale comune

I (documenti) e le query sono considerati come vettori incorporati in un spazio euclideo ad alta dimensionalità.

Spazio vettoriale
V è uno spazio vettoriale su un campo F (ad esempio il campo del reale o del complesso numeri) se:
un'addizione vettoriale di operazioni definita in V, indicata con v + w (dove v, w ∈ V)
un'operazione, moltiplicazione scalare in V, denotata a * v (dove v ∈ V e a ∈ F)
valgono le seguenti proprietà per ogni a, b ∈ F e u, v e w ∈ V:
v + w appartiene a V
u + (v + w) = (u + v) + w
Esiste un elemento neutro 0 in V, tale che per tutti gli elementi v in V, v + 0 = v

Per ogni v in V, esiste un elemento w in V, tale che v + w = 0
a* (b*v) = (ab)*v
Se 1 denota l'identità moltiplicativa del campo F, allora 1 * v = v
v + w = w + v
a * v appartiene a V
a* (v+w) = a*v+a*w
(a+b)*v=a*v+b*v

Corrispondenza parziale
classifica
Schemi di ponderazione

Distanza:
Possiamo calcolare la norma.
Norma L2 (norma euclidea):

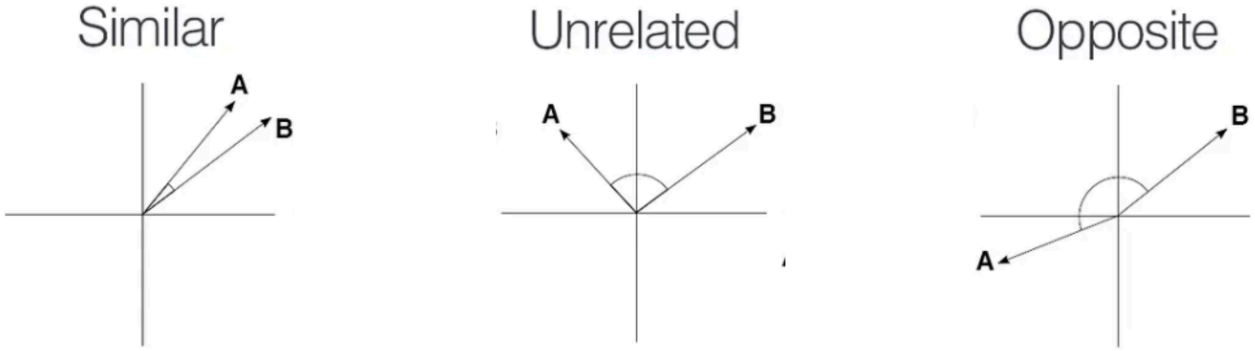
$x = \{x_1, x_2, x_3\}$

$y = \{y_1, y_2, y_3\}$

$x\text{-}y = (x_1 - y_1, x_2 - y_2, x_3 - y_3)$

L2-norm = $\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$

Somiglianza del coseno
A = $\{x1, x2, x3\}$ B = $\{y1, y2, y3\}$
È una metrica che misura il coseno dell'angolo compreso tra due vettori proiettato in uno spazio multidimensionale



$$\cos\theta = \frac{a \cdot b}{\|a\|_2 \|b\|_2}$$

-1 = vettori fortemente opposti → nessuna somiglianza
0 = vettori ortogonali → indipendenti
1 = sovrapposizione →elevata somiglianza

Distanza coseno:
Distanza coseno = 1 – somiglianza coseno

Se 2 vettori sono perfettamente uguali:
Somiglianza coseno = 1 (angolo = 0, $cos(\theta)=1$)
Distanza coseno = 1 – somiglianza coseno = 0

Bhattacharya Distanza
È una misura della somiglianza tra due distribuzioni di probabilità.
Ci dice quanta sovrapposizione c'è tra le due distribuzioni, e può aiutaci a determinare quanto sono simili o dissimili.

BD(P, Q)=-ln(BC(P, Q))

P, Q sono due distribuzioni di probabilità
BC(P,Q) è il coefficiente Bhattacharyya

$$BC(P,Q) = \sum(sqrt(p(x) * q(x)))$$

la somma viene presa su tutti i possibili risultati o eventi x
p(x), q(x) le probabilità di accadimento dell'evento x nelle distribuzioni P e Q

Misure di valutazione: Rilevanza binaria

Metriche insensibili agli ordini
Veri positivi
Veri negativi
Falsi positivi
Falsi negativi

Precisione:

$$Precision@k = \frac{true\ positives@k}{(true\ positives@k) + (false\ positives@k)}$$

k	1	2	3	4	5
Precision@k	$\frac{1}{1} = 1$	$\frac{1}{2} = 0.5$	$\frac{2}{3} = 0.67$	$\frac{2}{4} = 0.5$	$\frac{3}{5} = 0.6$

Non considera la posizione degli elementi rilevanti

Richiamare:
Indica quanti risultati effettivamente rilevanti sono stati mostrati tra tutti risultati effettivamente rilevanti per la query.

$$Recall@k = \frac{true\ positives@k}{(true\ positives@k) + (false\ negatives@k)}$$

k	1	2	3	4	5
Recall@k	$\frac{1}{(1+2)} = \frac{1}{3} = 0.33$	$\frac{1}{(1+2)} = \frac{1}{3} = 0.33$	$\frac{2}{(2+1)} = \frac{2}{3} = 0.67$	$\frac{2}{(2+1)} = \frac{2}{3} = 0.67$	$\frac{3}{(3+0)} = \frac{3}{3} = 1$

Punteggio F1:
Si tratta di una metrica combinata che incorpora sia Precision@k che

Recall@k prendendo la loro media armonica

$$F1@k = \frac{2 * (Precision@k) * (Recall@k)}{(Precision@k) + (Recall@k)}$$

k	1	2	3	4	5
Precision@k	1	1/2	2/3	1/2	3/5
Recall@k	1/3	1/3	2/3	2/3	1
F1@k	$\frac{2*1*(1/3)}{(1+1/3)} = 0.5$	$\frac{2*(1/2)*(1/3)}{(1/2+1/3)} = 0.4$	$\frac{2*(2/3)*(2/3)}{(2/3+2/3)} = 0.666$	$\frac{2*(1/2)*(2/3)}{(1/2+2/3)} = 0.571$	$\frac{2*(3/5)*1}{(3/5+1)} = 0.749$

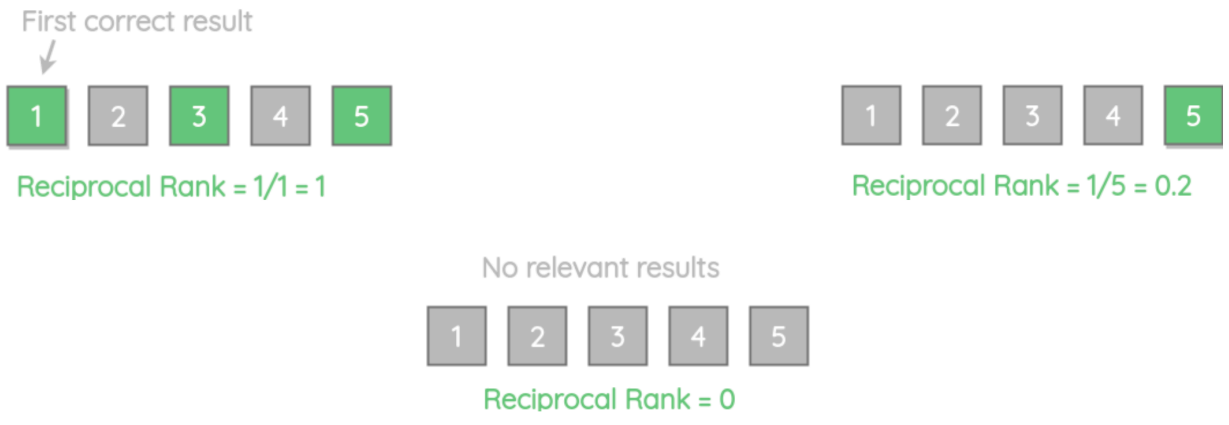
Metriche consapevoli degli ordini:

Grado reciproco medio (MRR)
Questa metrica è utile quando vogliamo che il nostro sistema restituisca il meglio elemento pertinente e desideri che quell'elemento sia in una posizione più alta.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

|Q|denota il numero totale di query
$rank1$ denota la classifica del primo risultato rilevante

Rango reciproco medio= il reciproco del rango del primo corretto rilevante risultato e il valore varia da 0 a 1.



Precisione media (AP)

Precisione media (AP)
È calcolata come la media delle precisioni per un certo numero di posizioni in questa lista

$$AP = \frac{\sum_{k=1}^{n}(P(k) * rel(k))}{number\ of\ relevant\ items}$$

rel(k) è una funzione indicatore che vale 1 quando l'elemento di rango K è rilevante
P(k) è la metrica Precision@k

Precisione media media (MAP)
È semplicemente la media della precisione media per tutte le query

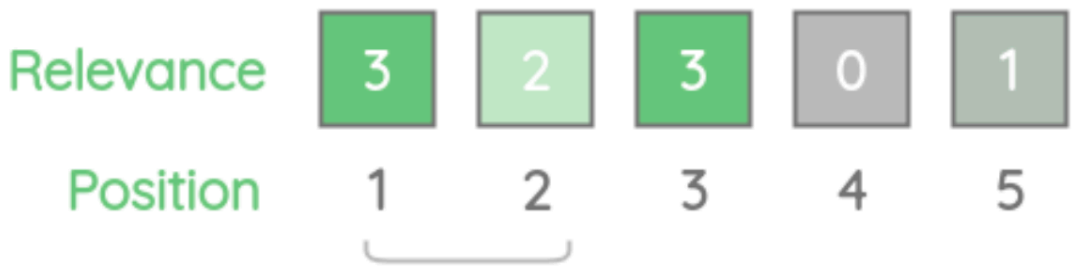
$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} AP(q)$$

Q è il numero totale di query
AP(q) è la precisione media per la query q

Guadagno cumulativo
Questa metrica utilizza un'idea semplice per riassumere semplicemente i punteggi di pertinenza
dei primi k elementi. Il punteggio totale è chiamato guadagno cumulativo.

Guadagno cumulativo scontato
Un elemento con un punteggio di pertinenza pari a 3 in posizione 1 è migliore dello stesso
elemento con punteggio di pertinenza 3 nella posizione 2.

$$CG@k = \sum_{1}^{k} rel_i$$



cumulative gain@2 = 3+2 = 5

Position(k)	1	2	3	4	5
Cumulative Gain@k	3	3+2=5	3+2+3=8	3+2+3+0=8	3+2+3+0+1=9

Guadagno cumulativo scontato
DCG introduce una funzione di penalità basata su log per ridurre la rilevanza punteggio in ogni posizione.

$$DCG@k = \sum_{i=1}^{k} \frac{rel_i}{log_2(i+1)}$$

Il guadagno cumulativo scontato prende semplicemente la somma di punteggio di pertinenza
normalizzato dalla penalità.

Normalizza i valori DCG utilizzando l'ordine ideale dei rilevanti elementi.

$$NDCG@k = \frac{DCG@k}{IDCG@k}$$