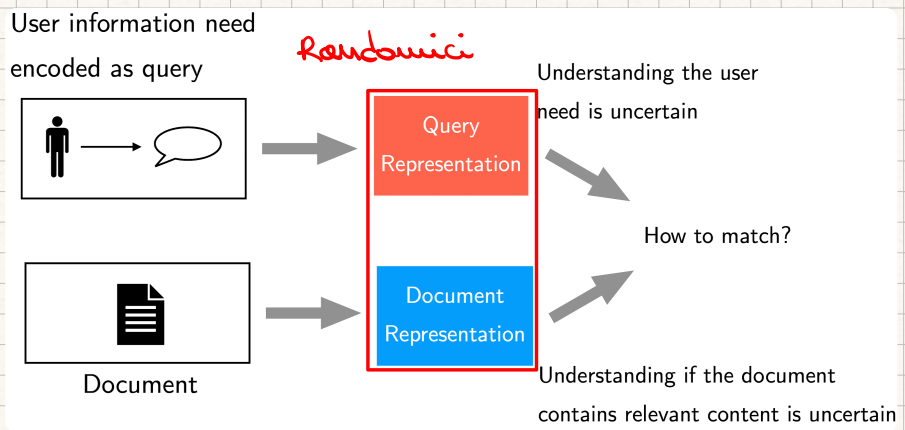
RANKED RETRIVAL

Fin ora abbiamo considerato solo query con risultati binari: o il documento fa match o no!
Però vorremmo anche capire, fra i vari documenti, l'utente non vuole scorrere migliaia di documenti che hanno termini in comune con la query, vuole vedere i migliori per poter scegliere, i top K.
Solitamente le query booleane ritornano o troppi o troppi pochi risultati, questo comporta molto tempo per processarle

Soluzione: Fare un rank dei documenti per la loro utilità, ordinarli dal più utile al meno utile e prendere i top-K.

Perché si usa la PROBABILITÀ?

Semplicemente perché la formulazione sia della query sia dei documenti non è DETERMINISTICA, anzi è piuttosto aleatoria.



PROBABILITY RANKING PRINCIPLE

Abbiamo una collezione di documenti D.
Abbiamo una query fatta da un utente q.
Vogliamo ritornare i top-k documenti.
Usiamo la seguente probabilità:

$$P(\text{relevance} \mid d, q)$$

Data la query e un singolo documento, con che probabilità esso è rilevante per la query?

Semplificazioni

1) Per il momento abbiamo un solo documento

2) Consideriamo per adesso la rilevanza come binaria.

Principio PRP

Se un sistema RDFORANCO RETRIEVAL risponde per ogni query con un rank dei documenti della collezione ordinati per la loro probabilità di utilità rispetto all'utente che ha fatto la query, dove questa probabilità è stimata al meglio possibile col bagaglio di tutti i dati disponibili al sistema, allora l'efficacia totale del sistema per l'utente è la migliore possibile considerando il bagaglio su quei dati.

migliore possibile → OTTIMA

Per utilità si intende la rilevanza del doc per la query.

IPOTESI

La relevance R è una variabile aleatoria che ha o valore 1 o 0. Bisogna fare un esperimento per capire se $R=1$ o $R=0$ dati documento e query

Notazione

$R=1 \rightsquigarrow r$

$R=0 \rightsquigarrow \bar{r}$

Anche la query è una variabile aleatoria con un valore generico $Q=q \rightsquigarrow q$

Il documento è una variabile aleatoria con un valore generico $D=d \rightsquigarrow d$

Possiamo ora definire:

$$P(R=1 \mid Q=q, D=d) = P(r \mid q, d)$$

Un sistema IR ritorna i top-K documenti $\langle d_1 \dots d_K \rangle$.
Diciamo che la **Recall** R_K è una variabile aleatoria che ci da indicazione sul **numero di documenti rilevanti** che il sistema ha ricavato, per una data query. Quindi $R_K \in [0, K]$.
D'altra parte abbiamo **l'efficacia totale** del sistema che viene misurata come il numero di documenti rilevanti **previsti** (Expected) che il sistema ritorna:

$$E[R_K]$$

DIMOSTRAZIONE DEL PRINCIPIO PRP

Teorema: con le adeguate assunzioni il PRP è valido!

Dimostrazione:

Bisogna provare che, $\forall K$ $\delta[R_K]$ è massimo se i documenti sono ordinati per la loro probabilità di rilevanza $P(r|q,d)$.

Definiamo $\delta[R_K]$ $\forall K$:

$$\delta[R_K] = r\sum_{i=1}^{K} P(r|d_i,q) + \bar{r}\sum_{i=1}^{K} P(\bar{r}|d_i,q)$$

Ma dato che la NON RILEVANZA $R=\emptyset$ $(\bar{r})$ possiamo omettere il secondo termine.

$$\delta[R_K] = r\sum_{i=1}^{K} P(r|d_i,q)$$

Dimostriamo **per contraddizione:**

Assumiamo esista un $\bar{K}$ tale per cui $\delta[R_K]$ è massima quando i documenti **non sono ordinati per rilevanza**.

Questo vuol dire che esiste almeno un documento d_l tale per cui:

$$P(r|d_l,q) < P(r|d_{\bar{K}},q)$$

Ovvero che ha una rilevanza minore dell'ultimo in classifica:

$$d_1 \dots d_l \dots d_{\bar{K}-1} \quad (2)$$

L'ordinamento naturale dovrebbe essere $d_1 \dots d_k$ (1), da cui diciamo che:

$$\delta[R_k|(2)] > \delta[R_k|(1)]$$

Ora, sviluppando:

$$\delta[R_k|(1)] = r \sum_{i=1}^{k} P(r|q, d_i) =$$

$$= r \sum_{i=1}^{k-1} P(r|q, d_i) + P(r|q, d_k)$$

$$>$$

$$r \sum_{i=1}^{k-1} P(r|q, d_i) + P(r|q, d_e) = \delta[R_k|(2)]$$

Questo è naturale perché, per definizione i k docs in (2) non sono ordinati, e $P(r|q, d_e) < P(r|q, d_k)$. Da qui la disuguaglianza che smentisce la contraddizione:

$$\delta[R_k|(1)] > \delta[R_k|(2)]$$

Per cui, con d_e, $\delta[R_k]$ non può essere massima.

ASSUNZIONI DEL PRP

1) Conosciamo $P(r|q,d)$ relativa a una sola query, un solo documento e una sola information need. Tutto per un singolo utente.

2) Assumiamo che l'efficacia del sistema sia $\delta[R_K]$. Ci aspettiamo di massimizzare anche la precisione a K:

$$\delta[P@K] = \frac{P[R_K]}{K}$$

Ma anche la recall a K:

$$\delta[R@K] = \frac{\delta[R_K]}{D_R} \longrightarrow \#\ \text{docs rilevanti per } q$$

PRP non è valida a priori per altre metriche

ESEMPIO DI COOPER

Assumiamo di avere due gruppi di utenti diversi, U_1 e U_2, tale che $|U_1| = 2|U_2|$.
Considerando una collezione di 10 documenti.
Ognuno dei due gruppi fa una query. Abbiamo che

1) $d_1 \dots d_9$ sono rilevanti per U_1
2) d_{10} è rilevante per U_2.

Da qui possiamo calcolare $P(r|q,d)$:

$$P(r|q,d) = \begin{cases} \frac{2}{3} & \text{per } d_1 \dots d_9 \\ \frac{1}{3} & \text{per } d_{10} \end{cases}$$

Questo perché d_{10} è rilevante per il doppio degli utenti presi per questo esperimento.

Se teniamo conto di due ordinamenti:

1) $d_1 \dots d_9\, d_{10}$, U_1 è felice, U_2 deve arrivare al decimo per trovare quello che cercava
2) $d_1\, d_{10} \dots d_9$, U_1 è felice col primo risultato e U_2 è felice col secondo

Quindi meglio il secondo ordinamento, anche se non segue il PRP. Qui PRP non vale perché abbiamo più di un singolo utente a fare la query.