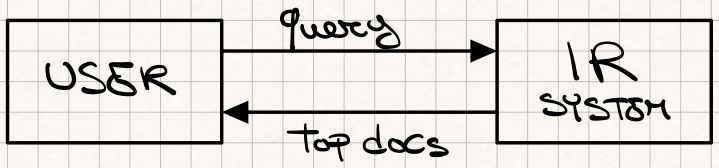
RELEVANCE FEEDBACK

I sistemi IR cercano di trovare ciò che in gergo è definita **information need** di un utente.



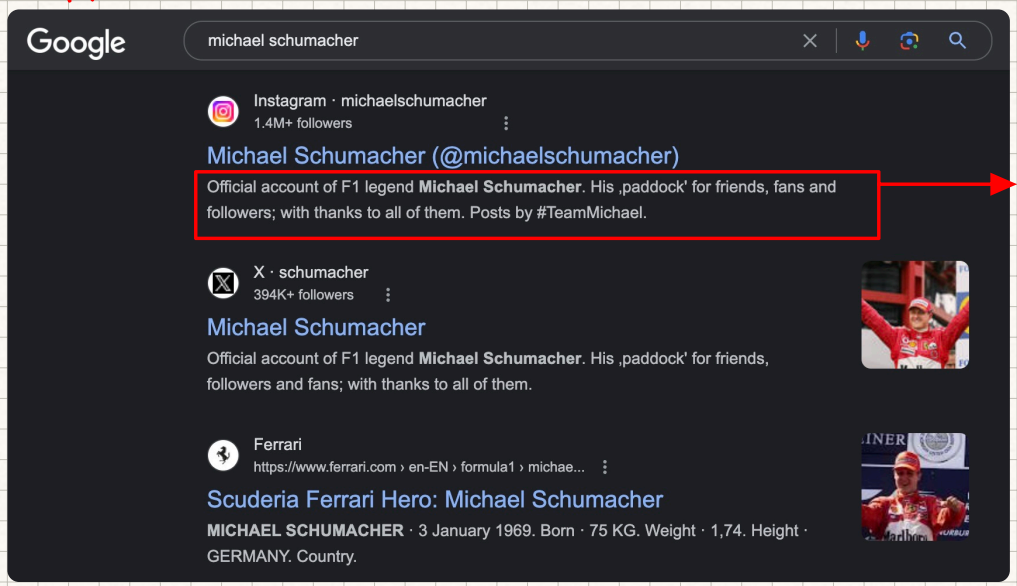
L'utente può volere una info esatta (Quando è nato Obama?) o una vaga (Problemi nell'acquistare un auto usata).

L'information need è spesso relata a un compito da svolgere (Come si risolvono i compiti di matematica?).

Ogni utente, data una information need, ha un modo diverso di scrivere la query.

Ambiguità: la stessa query può rappresentare information need diverse (Esempio di query: "Cooper", chi stai cercando? L'Attore, il cantante lo scienziato? Altro?)

SNIPPET



Questo è lo snippet

In base allo snippet l'utente si fa un'idea del contenuto di un documento.

USER INTERACTION

L'utente interagisce col sistema IR, in due momenti:

1) Quando formula la query
2) Quando consulta i risultati

L'utente può cambiare i risultati di un sistema IR tramite le sue interazioni con esso. Esistono diversi tipi di interazione:

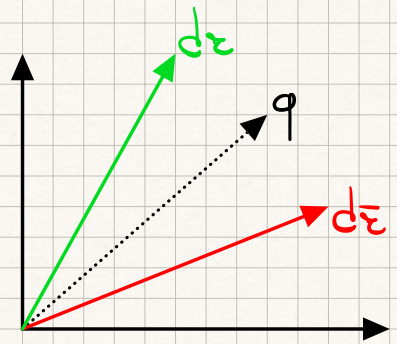
Explicit Interaction

IR sistema aiuta l'utente a perfezionare il risultato della query. Per farlo può chiedere un feedback sulla rilevanza di ogni documento ritornato, può

suggerire correzioni o integrazioni della query scritta dall'utente (Did you mean ...) oppure suggerire nuove query correlate a quella dell'utente.

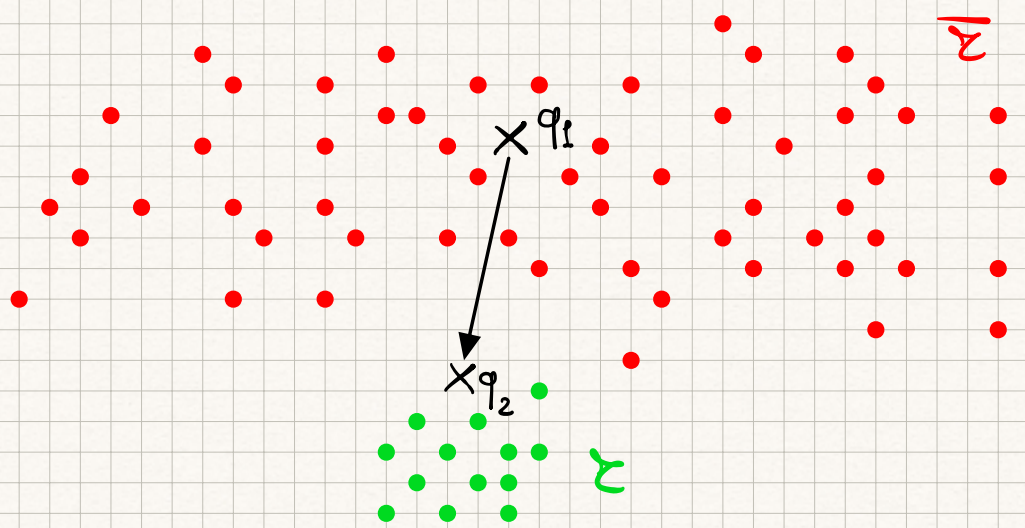
RELEVANCE FEEDBACK USING VECTOR

Dato che possiamo rappresentare, in uno spazio vettoriale sia la query che i documenti, ci aspettiamo che i docs rilevanti per una query siano vicini alla query stessa nello spazio.



d_r rilevante
$d_{\bar{r}}$ non rilevante

L'idea è quella di spostare la query verso i documenti rilevanti, nello spazio.



Per trasformare q_1 e portarla a q_2 bisogna togliere o aggiungere termini.
Solitamente si danno dei pesi positivi a quei termini che appaiono nei d_r, pesi negativi se il termine appare anche nei documenti $d_{\bar{r}}$. Sicuramente si eliminano i termini che appaiono solo in $d_{\bar{r}}$.

ALGORITMO DI ROCCHIO

Una query ottima massimizza la distanza fra il vettore medio dei d_r con quello dei $d_{\bar{r}}$:

$$\vec{q}_1 = \alpha \vec{q} + \beta \frac{1}{D^+} \sum_{d \in D^+} \vec{d} - \gamma \frac{1}{D^-} \sum_{d \in D^-} \vec{d}$$

Centroide dei d_r — Centroide dei $d_{\bar{r}}$

I parametri α, β e γ pesano i vari contributi, valori tipici sono $\alpha = 8$, $\beta = 16$ e $\gamma = 4$.

I termini della query che hanno un peso negativo, vengono buttati

Per evitare che la query cresca troppo, vengono aggiunti solo un numero ristretto di termini (tipicamente 50) fra quelli che hanno peso maggiore, quindi rilevanti.

PSEUDO RELEVANCE FEEDBACK

In questo approccio si tenta di usare il feedback degli utenti senza usare un input esplicito dell'utente. Come? Semplicemente assumendo che i top-K documenti ritornati siano rilevanti, e usarli per riformulare la query, con algoritmi tipo Rocchio.
Questo meccanismo è detto (automatic) query expansion. Nella query estesa non vengono rimodulati i pesi dei termini.

Problema del query-drift

Quando vengono usati documenti non rilevanti per riformulare la query, in particolar modo quando i top-K contengono pochi o nessun documento rilevante.