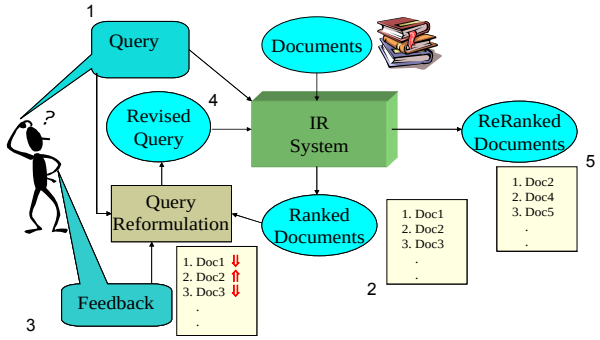
Relevance Feedback

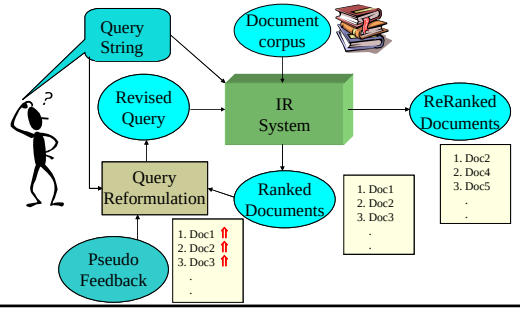
Relevance Feedback

- Dopo aver eseguito una query si chiede all'utente di selezionare i risultati che ritiene più rilevanti
- Si usa questo feedback per riformulare la query
- Si presentano nuovi risultati all'utente e, eventualmente, si re-itera il processo



Pseudo-Feedback

- Il feedback esplicito è poco usato
  - Soprattutto perché gli utenti sono spesso riluttanti
- Pseudo-Feedback: non chiede aiuto esplicito all'utente
  - Si assume che i primi *m* (top-ranked) documenti siano i più interessanti e si modifica la query di conseguenza



Riformulazione della Query

- Due tipi di approcci per riformulare una query:
  - Term Re-weighting: si aumenta il peso dei termini che compaiono nei documenti rilevanti e si diminuisce il peso di quelli che non vi compaiono.
  - Query Expansion: Si aggiungono alla query nuovi termini estratti dai documenti prescelti

Term re-weighting

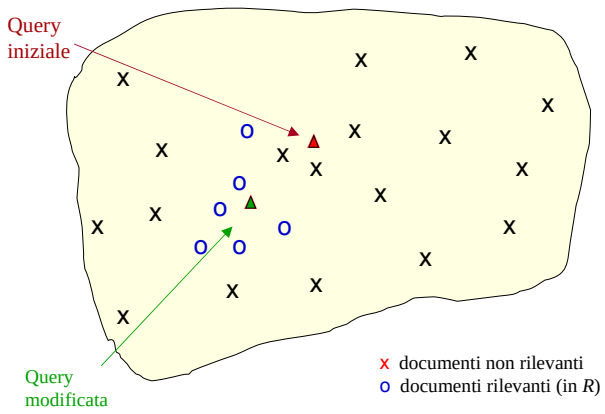
Term re-weighting nel modello vettoriale

- Si può modificare direttamente il vettore della query:
  - **Aggiungi** i vettori dei documenti **ritenuti rilevanti** (**feedback positivo**)
  - **Sottrai** i vettori dei documenti **ritenuti irrilevanti** (**feedback negativo**)
- Formula di Rocchio

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

  - q_m = vettore della query modificato
  - q_0 = vettore della query originario
  - D_r = documenti ritenuti rilevanti
  - D_{nr} = documenti ritenuti non rilevanti
  - α,β,γ: pesi (scelti manualmente o in modo empirico)
- Osservazione
  - Di solito si usa peso maggiore per il feedback positivo (es. γ=0.25, β=0.75).
  - Molti sistemi considerano solo feedback positivo (γ=0).

Term re-weighting nel modello vettoriale (algoritmo di Rocchio)



Term re-weighting nel modello probabilistico

$$Sim(d_j,q)=\sum_i w_{ij} w_{iq} \; [\; \log(\mathbf{P}(t_i|R)/(1-\mathbf{P}(t_i|R)) + \log((1-\mathbf{P}(t_i|\underline{R}))/\mathbf{P}(t_i|\underline{R})) \;]$$

- Dopo l'esecuzione della query
  - $\mathbf{D}_r$ = insieme di documenti restituiti (con ranking > una data soglia *r*)
  - $\mathbf{D}_{r\,i}$ = sottoinsieme di $\mathbf{D}_r$ contenente t_i
- Si può raffinare la stima delle probabilità
  - $P(t_k|R) = |\mathbf{D}_{rk}| / |\mathbf{D}_r|$
  - $P(t_k|\underline{R}) = (n_i - |\mathbf{D}_{rk}|) / (N - |\mathbf{D}_r|)$
- Aggiunta di fattori correttivi, per rendere la stima più robusta ad errori (dovuti a valori delle probabilità ~ 0)
  - $P(t_k|R) = |\mathbf{D}_{rk}| / |\mathbf{D}_r|$
    $\rightarrow P(t_k|R) = (|\mathbf{D}_{rk}| + n_i/N) / (|\mathbf{D}_r| + 1)$
  - $P(t_k|\underline{R}) = (n_i - |\mathbf{D}_{rk}|) / (N - |\mathbf{D}_r|)$
    $\rightarrow P(t_k|\underline{R}) = (n_i - |\mathbf{D}_{rk}| + n_i/N) / (N - |\mathbf{D}_r| + 1)$

Relevance Feedback sul Web

- Alcuni motori di ricerca offrono l'opzione *similar/related* pages
  - è una forma semplificata di relevance feedback
  - Esempi: Google, Altavista
- Relevance feedback per immagini
  - http://nayana.ece.ucsb.edu/imsearch/imsearch.html

Query expansion

Query expansion con Thesaurus

- Un thesaurus memorizza relazioni di sinonimia e associazione

```
physician
  syn: croaker, doc, doctor, MD, medical,
          mediciner, medico, sawbones
  rel: medic, general practitioner, surgeon,…
```

- Per ogni termine *t* in una query, si espande la query con i termini correlati a *t* (nel thesaurus)
  - In genere i pesi dei termini aggiunti sono più bassi
  - In genere questo metodo aumenta la recall, ma può diminuire la precisione (per via dell'ambiguità semantica)

Query expansion senza thesaurus: Automatic Global Analysis

- Automatic Global Analysis
  - Permette di valutare correlazioni fra termini, in assenza di un vero thesaurus
  - Calcola matrici associative che quantificano la correlazione fra termini, sfruttando statistiche basate sulla prossimità dei termini
- Uso per Query Expansion:
  - Per ogni termine i della query, espandi con gli *n* termini con i valori più alti di correlazione c_{ij}
- Problemi
  - Ambiguità: "Apple computer" → "Apple red fruit computer"
  - Poiché i termini sono, in ogni caso, altamente correlati, l'espansione potrebbe non aggiungere molti nuovi documenti rispetto alla query non espansa!

Query expansion senza thesaurus: Automatic Local Analysis

- Al momento della query, determina dinamicamente i termini simili usando i documenti top-ranked sulla base dei criteri classici
- L'analisi dei termini correlati non è basata sull'intera collezione, ma solo sui documenti "localmente" recuperati con la query iniziale
- Questo riduce il problema della ambiguità semantica:
  - i documenti, recuperati in base a tutti termini della query, molto probabilmente contengono ogni termine nel senso corretto per l'utente
  - "Apple computer" → "Apple computer Powerbook laptop"

Osservazione sui metodi di query expansion

- Local Analysis vs. Global Analysis
  - Nell'analisi globale i calcoli sono fatti una volta per tutte.
  - L'analisi locale deve essere fatta in tempo reale, per ogni query
    - Ha un aggravio computazionale continuo
    - ... ma fornisce risultati migliori
- Impatto sulla qualità dei risultati
  - L'espansione delle query può migliorare le prestazioni, in particolare la recall
  - Tuttavia, l'ambiguità semantica può influire negativamente sulla precisione
  - Metodi di WSD (word sense disambiguation) per selezionare il senso corretto