Sistemi Informativi
Laboratorio 4

Catalin Copil Mattia de Stefani Giulio Lovisotto

April 23, 2015

1 Descrizione

Visto che abbiamo scelto di usare BM25 per il ranking, applicheremo la formula tenendo in considerazione i giudizi di rilevanza. Per il relevance feedback esplicito utilizzeremo il file `qrels-originale.txt`. Ricordiamo che la formula di BM25 tieni gia' in considerazione i giudizi di rilevanza nella sua forma base.

$$\sum_{i \in Q} \log\Big(\frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)}\Big) \cdot \frac{(k_1 + 1)f_i}{k + f_i} \cdot \frac{(k_2 + 1)qf_i}{k_2 + qf_i}.$$

Ricordiamo che R e' il numero di documenti rilevanti per la query in questione, mentre r_i e' il numero di documenti rilevanti che contiene il termine i.

1.1 Relevance Feedback Esplicito

Il reperimento avvera' in 2 step. Nel primo verra' eseguito il ranking senza informazioni di rilevanza, e tra i primi N documenti verranno estratti quelli rilevanti usando il file `qrels-originale.txt`. Poi verranno estratti i valori R, r_i tra i documenti rilevanti individuati e verranno usati per la seconda esecuzione dell'algoritmo.

1.2 Pseudo Relevance Feedback

Il reperimento avvera' in 2 step. Nel primo verra' eseguito il ranking senza informazioni di rilevanza, verranno considerati i primi N documenti come rilevanti. Poi verranno estratti i valori R, r_i tra i documenti rilevanti individuati e verranno usati per la seconda esecuzione dell'algoritmo.

2 Implementazione

Per il calcolo di R ed r_i, utilizzeremo la matrice che contiene la frequenza di occorrenza delle parole per ogni documento (n_docs $\times$ n_words). Durante il reperimento, per ogni documento, per ogni termine i andremo a prendere il numero di documenti rilevanti che contiene il termine i, e lo salveremo in una mappa *map* (mappa $i \to r_i$). Useremo una funzione per il ranking che accetta in input i parametri R e la mappa *map* che mappa i termini i sul numero di documenti rilevanti che contiene i, e calcola il punteggio usando tali informazioni. I documenti rilevanti sono calcolati come descritto nella precedente sezione.