Analysis for Query Enrichment

For query Enrichment, we have selected Pseudo Relevance feedback technique. We have selected the top 10 documents fetched from BM25 as the relevant set of documents and the rest of the documents are considered as non-relevant. These values are selected empirically.

We have implemented Rocchio algorithm for query expansion. The formula is given as:

$$q'_j = \alpha.q_j + \beta.\frac{1}{|Rel|}\sum_{D_i \in Rel} d_{ij} - \gamma.\frac{1}{|Nonrel|}\sum_{D_i \in Nonrel} d_{ij}$$

In the text book the values for α, β and Δ are given as 8,16,4.We have modified these 3 parameters to 0.20,0.75,0.05 . These values were selected on the basis of a research paper - "Improving Retrieval Performance by Relevance Feedback." Gerard Salton; Chris Buckley. Journal of the American Society for Information Science (1986-1998); Jun 1990." We have tweaked the values even further and these changes were made empirically.

Following are the steps to implement the algorithm:

Step 1. Fetch documents with BM25 retrieval algorithm

Step 2. Select the top 10 documents as the relevant documents and count the term frequency for each term in these 10 documents

Step 3. Normalize this quantity by the root of summation of squares of frequencies of all the terms.

Step 4. Apart from the top 10 documents, rest of the documents are considered as non-relevant documents and the term frequency is calculated in the same way as step 2

Step 5. Above quantity is normalized in the same way as in step 3

Step 6. Rocchio's algorithm is used on every term in the corpus to generate a score for them.

Step 7. Top 20 scoring documents are inserted into the query

Step 8. Retrieval is done again using BM25 algorithm

Step 9. Steps 1 to 8 is repeated for all the queries.