Query Enrichment Approach- Pseudo relevance feedback

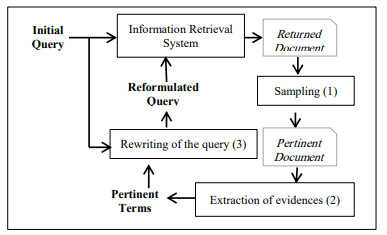
Pseudo relevance feedback (PRF) is a type of technique where the user is not prompted to identify the relevant documents; the system simply assumes that the top ranked documents are relevant. High frequency words from those are used to expand the query to fetch better results. The expansion terms generated by pseudo-relevance feedback will depend on the whole query, since they are extracted from documents ranked highly for that query, but the quality of the expansion will be determined by how many of the top-ranked documents in the initial ranking are in fact relevant.

Reasons for choosing PRF as our query enrichment technique:

* It doesn’t require user intervention.
* The words that occur more frequently in the relevant documents are used to expand the query which will result in getting better results.

Algorithm/Approach used:

* Pseudo relevance model was run on the BM25 scores for the queries as it had higher Mean Average Precision (MAP) value compared to other baselines.
* Initially a query Q was considered and the top K documents were obtained.
* The most commonly occurring words in those documents were considered after removing the stop words provided in the common\_words file.
* Different number of top k documents and top n frequently occurring combinations were considered and evaluated.
* These words were added to the query and this revised query is used to obtain the results.
* The steps involved can be represented as follows:



According to ‘Magdy W. and G. J. F. Jones. A Study on Query Expansion Methods for Patent Retrieval PAIR 2011 - CIKM 2011’ paper. It was stated that the best evaluations scores were found when 5 most frequently considered words were considered from top 5 documents

Trial 1:

Criteria- Three most frequently occurring words in top 5 documents were considered.(K=3,n=5)

* MAP: 0.358
* MRR: 0.572

Trial 2:

Criteria: Five most frequently occurring words in top 5 documents were considered(K=5,n=5)

* MAP: 0.424
* MRR: 0.673

The evaluation clearly backed the claim made in the paper and hence we considered the second trial’s approach in the project. Further increasing the number of words will broaden the search and may not do well against all queries.

Further improvement:

Words like 'cacm', 'jb', '1978' were added for majority of the queries. These might be considered as the stop words for the corpus and not included in the expansion.

Citation:

<http://doras.dcu.ie/16517/1/A_Study_on_Query_Expansion_Methods_for_Patent_Retrieval.pdf>

Snippet Generation

Successful interaction with the user depends on user’s understanding of the results. Thus many techniques have been used to display the results that are easier to understand.

Steps:

* Top ranked documents from BM-25 were considered for snippet generation.
* We obtained the sentences from the raw documents by splitting based on full stops. Only the valid sentences are considered.
* We have used the Luhn’s approach to rank each sentence in the document using a significant factor.
* The formula used to calculate the significant factor score for a sentence is:

Significance score = (number of significant words in the sentence)^2 / total number of words

* Here, the query words are considered as significant words. These significant words are emphasized using bold in the html files.
* The top 2 sentences in the document with the highest value of the score is taken and used to summarize the document.
* Blue color font is used to indicate the queries. The top ranked documents for it is indicated along with it’s summary. The figure below shows the snapshot of the snippet generation.

