Pseudo-relevance feedback

In general, relevance feedback means that user is involved in retrieval process. User can send a query, and the system gives back a set of results. After that, user sends a feedback about the relevant and non-relevant documents, then system process a better result based on the user’s feedback and show the final result to him/her.

You should consider that for this kind of technique, user has to have enough knowledge about the documents to be able to make an initial query. Also, there are some problems that this technique couldn’t solve it alone. For example: misspelling, cross-language information retrieval and mismatch of searcher's vocabulary versus collection vocabulary.

Misspellings: when user uses the wrong spell of a term that it is in the documents, relevance feedback is not useful.

Cross-language information retrieval: for documents in another language is not effective.

Mismatch of searcher’s vocabulary versus collection vocabulary: if the term that we search is different with the collection vocabulary, it will be not effective.

There are three types of feedback such as: explicit feedback, implicit feedback and blind or “pseudo” feedback. We are going to talk about pseudo-relevance feedback and we use in our project.

Pseudo-relevance feedback

In this technique, user sends a simple query and then user assume that top K files are relevant and he/she takes that. After that, user considers all these documents are relevant so, we do query expansion, and add these weighted terms from results to query. And finally return the most relevant documents.

There are several algorithms for doing this technique. We use Rocchio Algorithm for doing that. It implements relevance feedback in VSM (Vector Space Model). You can see the formula in below:

: Expanded or Optimal query

: Original query

: Set of relevant result documents (which is top k results in here.)

: Set of non-relevant result documents

: Weights