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 can see this kind of technique that is shown as “more like this” in some web pages.

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 that we use in our project.

Pseudo-relevance feedback

In this technique, user sends a simple query and then system assumes that top K files are relevant. 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 system returns the most relevant documents.

There are several algorithms for doing this technique. We use Rocchio Algorithm to do 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

Starting from, new query moves towards relevant documents and away from non-relevant documents. In this algorithm, negative term weight is ignored. It means that it is 0. So, the formula would be like this:

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.

The other problem is that most users are unwilling to send the feedback and make searching process longer. So, they prefer to just receive their result immediately. Also, in this technique, system should spend much more time to analyze documents, so queries that are longer are usually slower. Then, the cost of the retrieval systems will increase. But we can reduce this problem by reweighting certain prominent terms.