Information Retrieval: Practical 2

Hilary Term 2012

The aim of this practical session is to extend your document retrieval system from Practical 1, by implementing the Rocchio feedback mechanism (see the lecture entitled "Query Expansion"). You should use the same index and document collection as for Practical 1.

Feedback Mechanism

The feedback mechanism works by taking relevance judgements from the user, and repositioning the query vector using the relevant and non-relevant documents provided:

$$q_{\text{new}} = \alpha\ q_{\text{init}} + \beta \sum_{d \in D_r} \frac{d}{|d|} - \gamma \sum_{d \in D_n} \frac{d}{|d|}$$

D_r are the relevant documents provided by the user, and D_n the non-relevant documents; α, β and γ are parameters of the feedback mechanism which determine the importance of the original query, and the importance of the relevant and non-relevant documents provided. You should find values for the constants that make sense in this context, but the values do not need to be particularly optimised.

TF-IDF Weights

q_{init} is defined as for Practical 1, namely a vector formed by taking all the terms in the index, and assigning a weight of 1 or 0 to each index term depending on whether the term is in the query or not. The document vectors d in the Rocchio formula are also defined as for Practical 1, namely a vector of weights formed by taking each term in the index and weighting that term using TF $\times$ IDF. Note that TF, for the purpose of these practicals, is just the term frequency, i.e., the number of times the term appears in the document; there is no need to use a more complex expression involving logs. Similarly for IDF: this is just the inverse of the number of documents in which a term appears. The document lengths $|d|$ are given in the same file as for practical 1:
/usr/local/practicals/ir/practical1/data/doc_lengths.txt

Relevance Judgements

The relevance judgements to use for the feedback mechanism can be found here: /usr/local/practicals/ir/practical2/data/feedback.txt .
A "1" next to the document id means that the document is relevant, so belongs to the set D_r; and a "0" means that the document is non-relevant, so belongs to the set D_n.

Information Need

As for Practical 1, the information need is information relating to "financial instruments being traded on the American stock exchange". You should form a query from this description by taking all the words in the description which appear in the index.

Evaluation

For evaluation purposes, you should use the judgments in the following file: /usr/local/practicals/ir/practical2/data/relevant_nofback.txt .
These are the judgements from Practical 1, but with the documents in the file relevant.txt removed.
Compare your results using feedback, with your original no-feedback system, for both information needs (in both cases using relevant_nofback.txt for evaluation). Use your evaluation method from Practical 1 (i.e., precision and recall values at rank; and average precision). You should find that the feedback mechanism does not affect both information needs equally. Explain why.

Report

Your report should contain a short description of how you modified your original system to accommodate the feedback mechanism, and some results showing how the performance of your system has improved. **You should also explain why the documents given in the feedback judgements need to be removed from the evaluation file.**
You should arrive at the final practical session in week 8 with a completed assignment and report. The final practical session is intended as a sign-off session only, and you are expected to have completed the assignment before you arrive. The practical report does not carry any weight towards the end-of-term assignment, but counts towards the requirement that you achieve an overall pass in your practicals.

Summary of What You Need to Do

1. Modify your retrieval system from Practical 1 so that the query vector can be modified using the feedback mechanism;
2. Using the relevance judgements provided, and the α, β, γ parameters given earlier, obtain a new set of relevant documents for Information Need 1 from Practical 1;
3. Evaluate the new set of relevant documents as before, comparing your results with the original, no-feedback system;
4. Consider why the documents given in the feedback judgements need to be removed from the evaluation file;
5. Write a short report as instructed.