Pseudo-Relevance Feedback

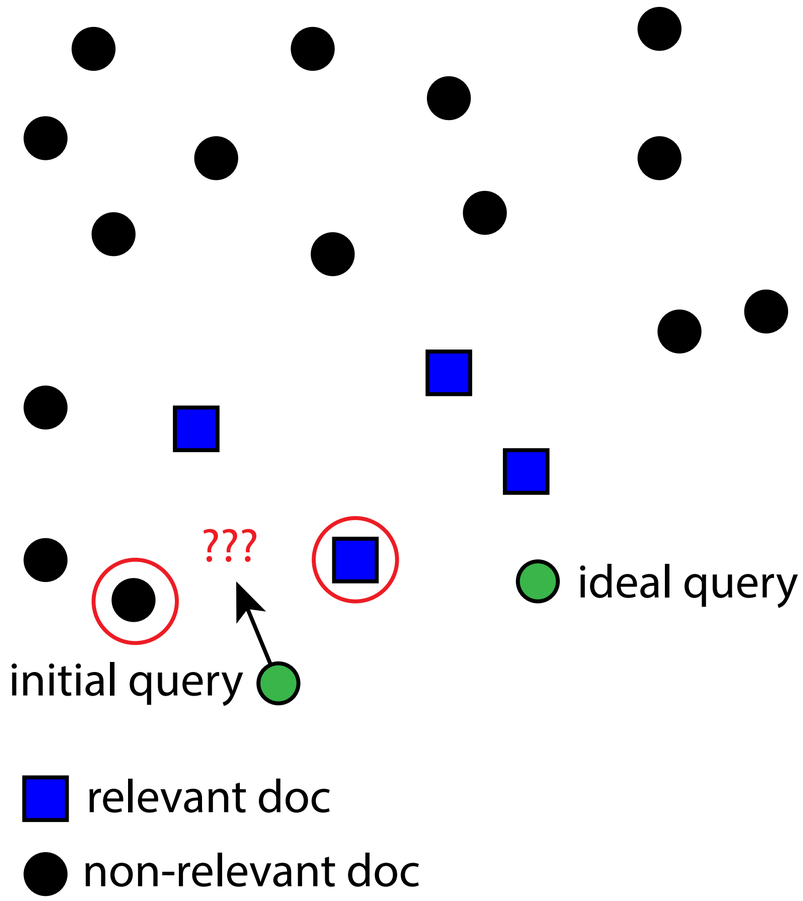
Pseudo-Relevance Feedback

If we assume the first k documents are relevant, we can update our query to find more relevant documents.

Rocchio's Algorithm for VSMs takes a linear combination of the original query and the set F of documents labeled as relevant:

$$\vec{q}_m = a\vec{q}_0 + \beta \frac{1}{|\mathcal{F}|} \sum_{\vec{d} \in \mathcal{F}} \vec{d}$$

How can we update this for language models?



Relevance Feedback with LM

A natural way to incorporate feedback documents into a query language model is to create a generative model of feedback documents, and smooth the query model together with it.

This generates an updated query model for use in Model Divergence Retrieval.

1. Generate query model $p(w|q)$.
2. Pick top k ranking documents as feedback set F.
3. Smooth query model together with feedback model, obtaining $p(w|q, F)$.
4. Rank documents using $p(w|q, F)$ as query model and display results.

Incorporating Feedback

One effective way to combine the query and feedback document models is to choose a model which minimizes average KL divergence between the query and feedback docs.

It's important to pay attention only to terms that are distinctive to the feedback documents in F, so we also want to maximize KL divergence to the corpus model C.

$$p(w|\mathcal{F},\mathcal{C}) := \arg\min_{\Theta} \frac{1}{\mathcal{F}} \sum_{i=1}^{|\mathcal{F}|} D_{KL}(\Theta \| p(w|\mathcal{F}_i)) - \lambda D_{KL}(\Theta \| p(w|\mathcal{C}))$$

$$\propto \exp\left(\frac{1}{1-\lambda} \frac{1}{|\mathcal{F}|} \sum_{i=1}^{|\mathcal{F}|} \log p(w|\mathcal{F}_i) - \frac{\lambda}{1-\lambda} \log p(w|\mathcal{C}) \right)$$

Feedback Model

$$p(w|q,\mathcal{F},\mathcal{C}) := a \cdot p(w|q) + (1-a) \cdot p(w|\mathcal{F},\mathcal{C})$$

Updated Query Model

Does it work?

This method consistently improves both average precision and recall. It finds more relevant documents, and places them higher in the ranking.

The disproportionate results from AP88-89 may be because vocabulary usage in this collection is more uniform, and thus easier.

		No Feedback	Feedback	Change
AP88-89	**AP**	0.21	0.295	40%
	Recall	3067/4805	3665/4805	19%
TREC8	**AP**	0.256	0.269	5%
	Recall	2853/4728	3129/4728	10%
WEB	**AP**	0.281	0.312	11%
	Recall	1755/2279	1798/2279	2%

Zhai et al, 2001

Comparing to Rocchio's Algorithm

Here we compare to Rocchio's algorithm using a VSM with BM25 term scores.

Average Precision has improved, but recall has decreased. This may be because the cutoff used to ignore low-probability words was more carefully tuned for the VSM.

For the LM approach, they calculate matching scores only for terms having $p(w|q, F) \geq 0.001$.

		Rocchio's	LM	Change
AP88-89	**AP**	0.291	0.295	1%
	Recall	3729/4805	3665/4805	-2%
TREC8	**AP**	0.26	0.269	3%
	Recall	3204/4728	3129/4728	-2%
WEB	**AP**	0.271	0.312	15%
	Recall	1826/2279	1798/2279	-2%

Zhai et al, 2001

Wrapping Up

This approach was developed in the following paper:

> Chengxiang Zhai and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In Proceedings of the tenth international conference on Information and knowledge management (CIKM '01), Henrique Paques, Ling Liu, and David Grossman (Eds.). ACM, New York, NY, USA, 403-410.

Pseudo-relevance feedback can make a big impact on retrieval performance, partly because queries tend to be under-specified. This approach, based on minimizing KL divergence, is just one possibility.