Pseudo Relevance Feedback: Report

Sidharth Agarwal 2019CS50661

October 11, 2021

Problem Statement

The goal is to develop "telescoping" model aimed at improving the precision of results using pseudo-relevance feedback. We will use data collection and queries from TREC COVID track. There are 192, 510 unique CORD identifiers in the collection, with each CORD id corresponding to one or more full text files that are extracted from article publisher website or from PubMed Central's (PMC) curated text site of the article.

Pseudo-Relevance Feedback with Rocchio's Method

Introduction:

The classical Rocchio's Method was introduced in Gerard Salton's SMART system in 1970s. It is built on the idea of query reformulation over the vector space model. It is aimed at finding a new query vector, which

- Maximizes similarity with relevant documents
- Minimizes similarity with non-relevant documents

In other words, reformulate the query such that, it gets closer to the neighborhood of the relevant documents in the vector space and it gets away from the neighborhood of non-relevant documents. The formula for Rocchio method is as stated below.

$$\overrightarrow{Q_m} = (a \cdot \overrightarrow{Q_o}) + \left(b \cdot \frac{1}{|D_r|} \cdot \sum_{\overrightarrow{D_j} \in D_r} \overrightarrow{D_j}\right) - \left(c \cdot \frac{1}{|D_{nr}|} \cdot \sum_{\overrightarrow{D_k} \in D_{nr}} \overrightarrow{D_k}\right)$$

Variable	Value
$\overrightarrow{Q_m}$	Modified Query Vector
$\overrightarrow{Q_o}$	Original Query Vector
$\overrightarrow{D_j}$	Related Document Vector
$\overrightarrow{D_k}$	Non-Related Document Vector
a	Original Query Weight
b	Related Documents Weight
c	Non-Related Documents Weight
D_r	Set of Related Documents
D_{nr}	Set of Non-Related Documents

Observations:

So the algorithmic details have been mentioned in the **algorithmic_details.pdf**. For reference the scores for the original ranking provided in *t40-top100.txt* were **nDCG = 0.7216 and MAP = 32.8818**.

I observed best performance by using the following factors for weights in different fields of xml.

Field	Factor of Multiplication
Title	4
Authors	2
Abstract	2
Body Text	1

I observed best performance with the *'query'* field of the xml queries. For fine tuning alpha, beta, gamma. Using the fact that $beta > gamma$, since positive feedback is more important than negative feedback.

alpha	beta	gamma	nDCG	MAP
0.25	0.75	0.15	0.702	31.010
0.50	0.50	0.15	0.718	32.731
0.75	0.25	0.15	0.731	33.094
0.75	0.50	0.15	0.718	32.828
0.9	0.2	0.15	0.732	32.740
0.90	0.25	0.15	0.731	33.015
1	0.25	0.15	0.730	32.963
1	0.25	0.25	0.703	29.966
1	0.50	0.15	0.721	32.873
1	0.75	0.15	0.715	32.604
1.25	0.25	0.15	0.728	32.822

1 Conclusion:

Since there are so many hyper parameters it is very likely that there is much more greater scope of improvement via fine tuning over my approach. But as per my observation the best results were obtained for $alpha = 0.75$, $beta = 0.25$ and $gamma = 0.15$, along with the above given table of weights over different fields of the document using the *'query'* field of the xml representation of the query.

References:

- Lecture Slides by Prof. Srikanta Bedathur, IIT Delhi
- Wikipedia Page on Rocchio's method - **link**