

Mini-Assignment 3: Query Expansion

On page 119, Baeza-Yates and Ribeiro-Neto give “the best query vector for distinguishing relevant documents from non-relevant documents” as $\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{\forall \vec{d}_j \in C_r} \vec{d}_j - \frac{1}{N - |C_r|} \sum_{\forall \vec{d}_j \notin C_r} \vec{d}_j$

where C_r is the set of documents relevant to the query and N is the size of the document set.

1. Does this equation represent expanding the query? reweighting the query terms? both?
2. Give two ways that C_r can be estimated.
3. The equation suggests that q_{opt} can include negative weightings for some terms. Is this appropriate in practice? Use an example to explain why or why not.
4. Does use of q_{opt} guarantee that the best document is ranked highest? If so give a proof; if not, give an explanation or a counterexample.
5. Explain how to modify the equation to decrease the possibility of query drift.
6. Is this equation adequate for finding explanatory web pages? Explain.
7. How would an ambiguous query term impact retrieval using this equation?
8. If the documents retrieved for a query are accurately clustered, query expansion (via pseudo-relevance feedback) can be done better. Explain how and rewrite the equation to model this.
9. Rewrite the equation to show how query expansion could be done using a thesaurus. Discuss the pros and cons of using this instead of pseudo-relevance feedback.
10. The equation as given handles (pseudo) relevance feedback that consists of binary judgments. Rewrite it to handle richer feedback.
11. Sketch out a caching scheme that can give users the benefits of query expansion with less delay, and outline the costs and benefits.
12. Computing the local relevance density of some terms in a document (or, similarly, looking for places where the stem-stem metric correlation is high) can be computationally costly. Explain how either of these, or a reasonable approximation, can be computed efficiently using the data structures seen earlier in the course.

Answer any 4 questions (if working individually) or 5 (if doing the assignment as a pair).

Due April 16