CS630 Lecture Notes: Problems

Lecturer: Lillian Lee
Scribes: Chris Danis (cgd3) & Brian Rogan (bcr6)

31 January 2006

In class we discussed two different types of document normalization functions. The normalization function that we covered extensively had the form:

$$norm_j^A(d) = norm^A(d) = \sqrt{(\Sigma_j \ tf_j(d)^2)}$$

We also briefly discussed using a function of the form:

$$norm^B(d) = \max_j(tf_j(d))$$

This problem will attempt to compare the use of different normalization functions, assuming we are using the standard inverse document frequency (idf) function. Assume that our corpus contains two documents, and they have the following term frequencies over the vocabulary:

	machine-learning	bayesian	network	conditionally-independent	bagging
$d^{(1)}$	14	12	1	1	1
$d^{(2)}$	8	7	7	2	3

1. Initially, it is quite clear that $norm^B$ will excessively punish documents that use one phrase often rather than related words. Thus, if documents use varied language to describe a concept, they are punished less than documents that use a smaller vocabulary to describe the same concept. Compute $norm^B$ for documents $d^{(1)}$ and $d^{(2)}$, and their overall ranking given the query $q =$"bayesian machine-learning" using a binary query vector (which is 1 if a term is in the query and 0 if a term is not). Use a term frequency function $tf_j(d)$ which is defined as the number of occurrences of vocabulary word v^j in document d . Which document does $norm^B$ suggest that we prefer?
ANSWER:
$norm^B(d^{(1)}) = 14$
$norm^B(d^{(2)}) = 8$

   Thus, for q, $score(d^{(1)}) = \frac{14+12}{14} = 1.85$ and $score(d^{(2)}) = \frac{8+7}{8} = 1.875$ so $norm^B$ suggests we should retrieve $d^{(2)}$.

2. Recompute the document ranking for $norm^A$. Does $norm^A$ suffer from the same bias described in the previous question? Does the ranking produced seem to accord with our intuitive idea of which document is more relevant?

   ANSWER:
   $norm^A(d^{(1)}) = 18.52$
   $norm^A(d^{(2)}) = 13.22$

   Thus, for q, $score(d^{(1)}) = \frac{14+12}{18.52} = 1.40$ and $score(d^{(2)}) = \frac{8+7}{13.22} = 1.11$, so we would prefer document $d^{(1)}$ for this query. The ranking that $norm^A$ produces does seem to accord with our idea of which document is more relevant: the subject of $d^{(1)}$ appears to be exactly our query. It is important to note, however, that in the general case $norm^A$ suffers from the same repetitive language bias as $norm^B$, just to a lesser extent. It is still the case that a document which uses one word many times, instead of several related words will be penalized more, because the square of the large value will dominate the squares of smaller values.

3. Singhal, Buckley and Mitra noted that for smaller documents, $norm^A$ overestimates relevance, and for larger documents it underestimates relevance. They call the point where $norm^A$ correctly estimates relevance the pivot point; however, this is not the only place on their graph where retrieval based on $norm^A$ and the true relevance intersect. There is a second point of intersection where the document length is roughly 20×10^3 bytes (see Figure 3). They suggest that the reason for this is that the cosine normalization function tends to behave like $(\#\ unique\ terms)^{.6}$ on TREC data. They claim though that retrieval is primarily affected by the number of terms in a query the document matches, so to cancel the bias of large documents, normalization functions that behave roughly like $(\#unique\ terms)$ should be used (basically, large documents have inflated relevance scores because their length gives them a greater chance of containing an element of the query). Is it likely that $norm^B$ is correlated with $(\#unique\ terms)$? Is it likely this correlation is linear? What does this say about the likeliness of $norm^B$ overestimating the relevance of large documents?

   ANSWER:
   It seems that $norm^B$ probably has some correlation with the number of unique terms, though this is likely quite accidental. This really all depends on the type of document. Repetitive documents that use one term in the vocabulary often will show a very weak correlation between $norm^B$ and $(\#unique\ terms)$, because normalization is based only on the most used word. If the author seeks to avoid repetition the correlation will likely be stronger, but really $norm^B$ is more likely to be correlated with document length: the author will use the most used phrase in proportion to the length of the document. There is certainly no evidence that $norm^B$

is linearly related to ($\#unique\ terms$); in fact it seems that it is more likely to be sub-linear than linear. If this is the case, then $norm^B$ will overestimate relevance of large documents. It would take a collection of documents where the most-used term is used more often than the number of distinct terms in the document. The authors note that "it is well known that majority of terms in a document occur only once", so this means that the most used-term would have to be used more often than almost all of the other terms in the document combined. This seems highly dubious.