CS276: PA3 report

Raj Bandyopadhyay, Rafael Guerrero

May 18, 2013

Introduction

In this assignment, we use different ranking algorithms to rank a pre-determined set of queries and their corresponding results. The parameters for tuning the ranking algorithms are selected manually.

Task 1: Cosine Similarity

In our implementation of cosine similarity, we use *sublinear scaling* for the document term vectors in order to reduce the effect of outliers. For *length normalization*, we assume a body length of 1000 where the length is not provided. The same body length is assumed for length normalization of the remaining (non-body) fields.

Weight choices

The best parameters that we obtained by hand-tuning the cosine similarity weights are as shown in in Table 1. The NDCG score obtained is 0.835. We did find other weight values that increased the NDCG, but we decided to select these values since we wanted to avoid *overfitting* to the training data.

Table 1: Weights for cosine similarity (NDCG 0.835)

Field	Weight
task1_W_url	-1.0
task1_W_title	-0.90
task1_W_header	-0.80
task1_W_body	-0.90
task1_W_anchor	-0.10

Effect of different weight choices

In the case of cosine similarity, the weights are negative because the use of log-scaling for the term frequency vectors and the length normalization constant leads to negative term frequency vector entries. However, in terms of magnitudes, we found that URL is the most important in determining the relevance of the page, followed by the title and the body. The anchors are the least important.

On experimentation, we found that the weights for the URL and the anchors had the most effect on the NDCG score, but in opposite ways. Increasing the magnitude of URL weight increases NDCG, while producing the opposite effect for anchors. This is likely because the URL has a strong signal while anchors have a lot of noise.

Task 2: BM25F

Table 2 shows the chosen weights for the BM25F parameters, which give us an NDCG value of 0.856. For the remaining parameters, we chose the following values:

- $K_1 = 1.0$
- $\lambda = 3.0$
- V_j : saturation function

Table 2: Weights for BM25F (NDCG = 0.856

Field	Weight	Field	Weight
$task2\_W\_url$	1	$task2\_B\_url$	1.0
$task2\_W\_title$	0.9	$task2\_B\_title$	0.1
$task2\_W\_header$	0.8	$task2\_B\_header$	1.0
$task2\_W\_body$	0.9	$task2\_B\_body$	1.0
$task2\_W\_anchor$	0.7	$task2\_B\_anchor$	0.1

Choice of weights

As we can see from Table 2, the fields are weighted in similar order of importance as in cosine similarity. The URL gets maximum weight while the anchor gets the lowest weight. The B parameters reflect a similar importance of the URL and anchor fields, while the other fields did not make a significant difference.

Effect of λ and K_1 parameters

We found that $K_1 = 1.0$ was the best value for K_1, after trying several values in the range of 0.5 to 2.0. We tried different values of λ in the range of 1.0 to 25. It made a very small difference (0.5%) to the NDCG score, so we decided to use the best value in that range i.e. 3.0.

Effect of V_j

We tried three $V_j$functions: saturation, sigmoid and logarithmic. On the whole, they did not have a huge effect. Table 3 shows the best NDCG values we obtained for the three functions. On the whole, we found very minor differences in the effect of the three functions. We decided to use the logarithmic function, because even though the saturation function gave better results in some cases, we think the logarithmic function is more stable.

Table 3: NDCG values for different V_j

V_j	NDCG
saturation	0.8564
sigmoid	0.8543
logarithmic	0.8562

Task 3: Smallest Window

We modified the code in task 1 (cosine similarity) to incorporate the smallest window metric. Our boost factor algorithm follows the assignment description. We use a value of $B = 2$. For values of $w_{q,d}$between query length $|Q|$ and ∞, we tested two methods:

1. Inverse boost function = $B\frac{|Q|}{w_{q,d}}$
2. Sigmoid boost function= $2 - \frac{1}{1+e^{-(w_{q,d}-|Q|)}}$

The cosine similarity weights were set to the same values as task 1 (Table 1), giving us an NDCG value of 0.857.

Effect varying B and Boost function

For Inverse boost function varying the B value allowed us to get a maximum improvement of 0.52%. The improvement we observed ranged from +0.06% for B=1024 to +0.52% for B=16. After reaching B=16 we didn't get any better NDCG value by increasing B. For Sigmoid Function varying the B value allowed us to get a max improvement of 0.23%. The improvement we observed ranged from +0.07% for B=2 to B=0.23% for B=16. After reaching B=16 we didn't get any better NDCG value by increasing B. This may be due to the fact that after a certain value of B, documents get a high boost based on fields that are not very relevant.

Other Questions (Page 10 of assignment)

Q1) Reasoning behind weights and properties of documents: A query term occurring in the URL or the title is an extremely strong signal of relevance since it indicates that the document is likely to be related to the user's information need, so we gave those the highest weights. The anchors get the least weight because the same anchor term can point to many different pages and thus provide a very noisy signal.

The main useful property of the documents we noticed in our training set was that the URLs are particularly well-formed. Once we parse the URLs, many of the resulting tokens are actual query or dictionary terms, making them a strong signal.

Q2) Other metrics that can be used for better scoring function:

- Age of the document (i.e. time of creation or last modification): More recent documents should be given higher priority.
- The frequencies of other (non-query) terms in the document: This is because some of those terms might be synonyms of the query terms or simply highly related (through lemmatization, for example).

Q3) BM25F parameters: See section on Task 2.

Q4) $V_j$function in BM25F: See section on Task 2.

Q5) Effect of varying B and Boost function in small window: See section on Task 3.