Homework 4

1 Introduction

1.1 Collaboration and Originality

1. Did you receive help of any kind from anyone in developing your software for this assignment (Yes or No)? It is not necessary to describe discussions with the instructor or TAs.

   No.

2. Did you give help of any kind to anyone in developing their software for this assignment (Yes or No)?

   No.

3. Did you examine anyone else's software for this assignment (Yes or No)? It is not necessary to mention software provided by the instructor.

   No.

4. Are you the author of every line of source code submitted for this assignment (Yes or No)? It is not necessary to mention software provided by the instructor.

   Yes.

5. Are you the author of every word of your report (Yes or No)?

   Yes.

2 Experiment 1: Baselines

2.1 Experimental Results

	BM25 (Exp-1a)	Indri BOW (Exp-1b)	Indri SDM (Exp-1c)
P@10	0.436	0.432	0.46
P@20	0.422	0.432	0.456
P@30	0.4173	0.4373	0.4547
NDCG@10	0.3481	0.3276	0.3475
NDCG@20	0.3509	0.3379	0.3559
NDCG@30	0.3538	0.3515	0.3649
MAP	0.2412	0.2576	0.2718

2.2 Parameters

retrievalAlgorithm=Indri

BM25:k_1=1.2

BM25:b=0.75

BM25:k_3=0

Indri:mu=2500

Indri:lambda=0.4

For sequence dependency, I choose 0.8 for #AND operator, 0.1 for #NEAR operator and 0.1 for #WINDOW operator.

3 Custom Features

Feature 17, the date of the document is collected. This feature could be very helpful, because people might be more interested in new document than old document. The computation complexity is very low.

Feature 18 finds out all the positions of terms in the query and calculates its standard deviation. The hypothesis here is that if all terms appear together rather than diversely in the document, the document could be a better match. Because terms could be use together to form some concept appearing in the query. I use the standard deviation of all the positions to represent how closely terms are used in the document. The time complexity of creating this feature is in the order of the length of the document. Because a pass of scan of all positions is needed.

4 Experiment 2: Learning to Rank

4.1 Experiment Table

	BM25	IR Fusion (Exp-3a)	Content-Based (Exp-3b)	Base (Exp-3c)	All (Exp-3d)
P@10	0.436	0.436	0.432	0.488	0.476
P@20	0.422	0.43	0.424	0.474	0.474
P@30	0.4173	0.4347	0.432	0.448	0.4587
NDCG@10	0.3481	0.3643	0.3574	0.3916	0.3913
NDCG@20	0.3509	0.3653	0.3594	0.3913	0.388
NDCG@30	0.3538	0.3717	0.363	0.3811	0.3885
MAP	0.2412	0.2517	0.2511	0.2595	0.2584

4.2 Parameters

retrievalAlgorithm=LeToR

BM25:k_1=1.2

BM25:b=0.75

BM25:k_3=0

Indri:mu=2500

Indri:lambda=0.4

4.3 Discussion

Overall

all re-rank system has a higher MAP score than the basic system, which show the effectiveness of the rer-rank algorithm.

Comparing 3a and 3b, adding more features does not necessary improve the performance. Other than IR Fusion (3a) features, 3b add overlap score. Since the BM25 and Indri model already use term matching internally in their algorithm, the overlap score could be a redundant information to use. Moreover, this score could be noisy, because it is very easy to have term overlapping but they do not actually have the same semantic. In Matching Model (Indri or BM25), the nosiness of the simple overlapping could be smoothed by balancing all kind of information. (the term might not only need to overlap, but only within a window or has close distance etc.). Therefore, using overlapping score in the final re-rank round could be harmful.

Base system (3b) has large improvement comparing to 2a, 2b. This shows the effectiveness of Spam score, Url depth, Wikipedia and PageRank feature. Moreover, we can see that the score of P@10 and NDCG@10 is improved more than P@30 and NDCG@30. This shows the information from Pagerank, spam score etc could help to find the top match for the query a lot. This is reasonable, because these features not only look at the text of query and document only, but also consider the quality of the document in the network too. Documents endorsed by high Pagerank, low Spamscore is likely to be the things that user is looking for. And these are very robust document prior.

Customize features 17, 18 are not very useful. Feature 17, which tries to use the date (time) information of a document, could be hard to modeled. For example, high quality document from long time ago could be a good match, whereas the latest documents which suits the trend could be a good match too. This means the relation of time and matching degree is not in simple linear relationship and can be hard for SVM to learn. Feature 18 tries to look at the concentration degree of query terms in the document. This feature might not be so useful. Because a good matching document could layout the related query terms in arbitrary position, thus the concentration degree (I use deviation) could very a lot.

Example

Many queries have worse performance after using feature 17 and 18. But query 31 goes from 0. 3782 to 0. 3920 after adding feature 17, 18. The query term is “Atari”. The most improvement happens in the top matching. P@10, P@15, P@20 all reach to 1.000 from 0.9x. As we can see the query is very vague and it is very hard to tell what the intention after the single word “Atari” is. At this time, it might be very useful to return a more recent document than an old document. It is possible there is lately some trend on “Atari” and the user is look for that. In the model, feature 17 actually has a weight of 0.49655315, which is pretty large and could be the decisive factor to decide which document to rank the top.

5 Experiment 3: Feature Combinations

5.1 Experiment Table

	All (Baseline) (Exp-4a)	$Comb_1$ (Exp-4b)	$Comb_2$ (Exp-4c)	$Comb_3$ (Exp-4d)	$Comb_4$ (Exp-4e)
P@10	0.476	0.508	0.464	0.492	0.504
P@20	0.474	0.466	0.47	0.46	0.464
P@30	0.4587	0.448	0.46	0.4547	0.452
NDCG@10	0.3913	0.4036	0.3879	0.3926	0.3954
NDCG@20	0.388	0.3906	0.3871	0.3841	0.3861
NDCG@30	0.3885	0.3845	0.3903	0.3869	0.3845
MAP	0.2584	0.2622	0.2586	0.2611	0.2611

5.2 Parameters

retrievalAlgorithm=Indri

BM25:k_1=1.2

BM25:b=0.75

BM25:k_3=0

Indri:mu=2500

Indri:lambda=0.4

Comb1 disabled feature: 7,10,13,16,17,18

Comb2 disabled feature: 4,12,14,15

5.3 Discussion

Comb1: According to Exp3c and Exp3d, content feature 7, 10, 13, 16 might not be so useful. And my custom feature 17 18 might not be so useful. So, I deactivate them all.

The result show large improvement. And the result is higher than the base Model, which use all features but custom features. This means that overlap score is not useful. And this proves my discussion in the 4.3.

Comb2: From the model of Exp-3d, all. I identify the features with low weight. They are feature 4, 12, 14, 15. Their weight in absolute value is less than 0.1. And I deactivate them to see what will happen. They are PageRank, BM25 and Indri score on inlink field is not so useful. But the overlap score on inlink is somewhat useful. This is reasonable. Because inlink can be simply view as Bag of words. If the query term appears, it could be a good evidence to match this document. And we probably do not need to have complicate matching algorithm like BM25 and Indri to do this on inlink.

The result does not improve a lot from the 4a. This is reasonable. Because these features originally have low weight in the system, which means they are not so important. Therefore, removing them might not have large effect on the model. And the results show that score on

Comb3: Inspired by Comb2, removing low weight feature might not harm the performance and might be help a little. I went to check the model output of Comb1 to see what other features I can eliminate. Using less feature could help to reduce the time and space complexity of model. Therefore, it is good. Feature 3, 4, 15 is removed on top of Comb1.

The result shows performance decreased. It is possible that a feature has a low weight because it has strong correlation with other features. But the feature itself is very useful. Feature 3 has weight of 0.084 in Exp-4b. And it is Wikipedia score. Feature 4 has 0.084 score too and it is PageRank score. These two scores are good document prior. It is very likely that they correlate with other features. Good quality document (high PageRank, Wiki score) has high BM25 and Indri scores too. These two features could be useful in some other cases when BM25 and Indri score are not good indicators. So, we should keep them. Feature 15 is inlink Indri score. Other features have similar information with this one. So, we should be able to remove it.

Comb4: Based on previous discussion. Add Pagerank and Wikipedia score back to the model.

The overall MAP score remains the same. But interestingly, the top matching score like P@10 and NDCG@10 improves. This is a strong evidence to show that PageRank and wikipeida score are very useful feature to pick out the top matching documents. Because these two score shows that a document is of high quality, useful and authoritative.

6 Analysis

Exp	Weight
4a	1 1:0.98496336 2:-0.27127877 3:-0.14518802 4:0.099152774 5:1.37958 6:0.27084592 7:0.56069791 8:0.25633079 9:-0.30194399 10:0.48823497 11:0.1633682 12:-0.065367475 13:0.32335016 14:0.010732621 15:0.043158088 16:0.13839906 17:0.49655315 18:0.47804293
4b	1 1:0.97812891 2:-0.25669071 3:0.084187403 4:0.085749462 5:1.5497392 6:0.33844644 8:0.79268265 9:-0.18493269 11:0.41470495 12:0.052929752 14:0.15783076 15:0.040438477
4c	1 1:0.93700254 2:-0.29010975 3:-0.16026941 5:1.3952957 6:0.29395026 7:0.57518828 8:0.27846521 9:-0.29373202 10:0.51188087 11:0.12597652 13:0.28525257 16:0.14235845 17:0.50485605 18:0.49025115
4d	1 1:1.0009145 2:-0.25173372 5:1.6033425 6:0.39194137 8:0.77741164 9:-0.19222492 11:0.42057693 12:0.058702555 14:0.16067217 #
4e	1 1:0.95454758 2:-0.23869106 3:0.083345644 4:0.065714814 5:1.5955545 6:0.37384143 8:0.73508632 9:-0.21597479 11:0.43725076 12:0.074425332 14:0.14001536 #

1. Removing low weight does not affect the performance of system that much. And I tried to remove 4,12,14,15 in Comb2(Exp-4c) experiment. This shows that Feature like Indri, BM25 score on inlike field is not so useful.

2. High weight does not necessarily mean the features is very usefully. And removing a high weight feature could actually improve the performance. In the Comb1 (Exp-4b), I remove my custom feature and also the overlap score feature on body field and title field. They actually have a large weight (check 4a) but removing them give a large improvement. The effect of overlap is largely captured by BM25 and Indri score, so adding them might not be so useful.

3. Some low weight score feature could be important, and we should not remove them. In the 4d and 4e. I tried to remove PageRank and Wikipidea score. They have low weight and removing them indeed doesn't damage my overall MAP score. But they are good document prior to use and they are actually very useful for the matching in the top range.