Your Name: Wei Wang

Your Andrew ID: ww5

Homework 4

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.

   If you answered Yes, provide the name(s) of anyone who provided help, and describe the type of help that you received.
   No

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

   If you answered Yes, provide the name(s) of anyone that you helped, and describe the type of help that you provided.

   No

3. 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.

   If you answered No:
    a. identify the software that you did not write,
    b. explain where it came from, and
    c. explain why you used it.

   Yes

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

   If you answered No:
    a. identify the text that you did not write,
    b. explain where it came from, and
    c. explain why you used it.

   Yes

Your Name: Wei Wang

Your Andrew ID: ww5

Homework 4

1 Experiment: Baselines

	BM25	Indri BOW	Indri SDM
P@10	0.1680	0.1720	0.1920
P@20	0.2040	0.1980	0.1960
P@30	0.2053	0.1893	0.2000
MAP	0.0924	0.1146	0.1183

BM25:k_1=1.2; BM25:b=0.75; BM25:k_3=0
Indri:mu=2500; Indri:lambda=0.4
fb=true; fbDocs=20; fbTerms=10; fbMu=0; fbOrigWeight=0.5
fbExpansionQueryFile=OUTPUT_DIR/HW4-Exp1-3.qry
fbInitialRankingFile=TEST_DIR/HW4-Exp1-2.fbRank

The BM25 and Indri are more effectiveness, it took around 20mins to run Indri SDM, and the result is showing a little improvement. This can be caused due to the number of document I set, since in my previous experiment, the best result always occur at around fbDocs=40, while it takes too much time.

2 Custom Features

My two custom features are document body length and document title length. The information it utilized is the field information as “body” and “title”, document id. My intuition is that the longer the field in this document, the more important this feature is. The computational complexity is O(N), since I only need to get the length of the field. By implementing in the following experiment, we can see the inclusion of these two features can help with the accuracy. Because for most of the document, the longer the feature is, the more information it puts into the part.

3 Experiment: Learning to Rank

	IR Fusion	Content-Based	Base	All
P@10	0.2240	0.2040	0.3440	0.3680
P@20	0.2280	0.2180	0.2860	0.3000
P@30	0.2067	0.2107	0.2760	0.2720
MAP	0.0993	0.1003	0.1235	0.1298

3.1 Parameters

BM25:k_1=1.2
BM25:b=0.75
BM25:k_3=0
Indri:mu=2500
Indri:lambda=0.4
letor:trainingQrelsFile=TEST_DIR/HW4-train.qrel
letor:trainingQueryFile=TEST_DIR/HW4-train.qry

These are the parameters I fixed in the learning-to-rank model. Qrels and Query are provided through the write up.

3.2 Discussion

The learning to rank system have a better result than the baseline model. Especially the performance are improved a lot when including two custom features.

When using all features, the system performs the best. The content-based even perform worse than IR Fusion. The "base" performance is closed to all features; therefore, this indicates that the two custom features can contribute to the accuracy. Compared content-based and IR Fusion, the differences are feature 7, 10, 13, 16, which are the term overlap score for feature body, title, url and inlink. This check to what percentage the term in query is overlap with the document field part. When adding these parts, the result is even worse, this indicates that the term overlap part not support to improve accuracy, maybe the doc has higher overlap rate, which means that it contains most of the word, is not the relevant document that the user wants. In addition, compared to base and all scenario, including feature 1-4 greatly improve the result. This can be caused due to what spam and page rank contribute to the result, these features have a high correspondent with the relevant score. I can later exam them in the following experiment of feature combinations.

4 Experiment: Features

Experiment with four different combinations of features.

	All (Baseline)	**Comb$_1$**	**Comb$_2$**	**Comb$_3$**	**Comb$_4$**
P@10	0.3680	0.2200	0.2320	0.3760	0.4000
P@20	0.3000	0.2220	0.2420	0.3140	0.3220
P@30	0.2720	0.2160	0.2320	0.2800	0.2827
MAP	0.1298	0.0921	0.1074	0.1361	0.1399

4.1 Parameters

BM25:k_1=1.2
BM25:b=0.75

BM25:k_3=0
Indri:mu=2500
Indri:lambda=0.4

Comb1: keep 5,6,14,15, they are the BM25 and Indri score for body and inlink
Comb2: keep 5,6,17,18, BM25 and Indri score for body, and document body length, document title length
Comb3: keep 1,4,5,17, spam, pagerank, BM25 for body, document body length
Comb4: keep 1,5,10,17, spam, BM25 for body, term overlap for title, and document body length
letor:trainingQrelsFile=TEST_DIR/HW4-train.qrel
letor:trainingQueryFile=TEST_DIR/HW4-train.qry

4.2 Discussion

There are 5 scenarios in this section. Except the first one include all features, the rest of them include 4 features as combination. Firstly I tried feature 5,6,14,15, they are the BM25 and Indri score for body and inlink, I set this scenario since the BM25 score and Indri score can be internal correlated, therefore, they might have a negative impact to accuracy. The result shows that it is relatively close to the IR Fusion result in the previous experiment part, which indicates that the IR Fusion don’t include much features that can work efficiently. Secondly I tried feature 5,6,17,18, BM25 and Indri score for body, and document body length, document title length, by removing 14, 15, and adding two custom features, I would like to see to what extent that two custom features can improve accuracy. And the result shows that it improves the accuracy to some extent. For example, the p@10 by 5.45%. Up to this point, there are two conclusions I think can help with the following design: 1. Don’t include both BM25 and Indri, they are internal correlated. 2. Feature 17 and 18 are helpful, while they might be correlated, so considering only include one of them.

Later in combination 3, I tried feature 1,4,5,17, which represent spam, pagerank, BM25 for body, document body length. This improve the accuracy significantly. In combination 4, I include 1,5,10,17, which are the spam, BM25 for body, term overlap for title, and document body length, and this get a better performance by improving p@10 by 8.7%, compare to the baseline. In a word, by adjusting the feature combination, the precisions and MAP are going to the right direction

5 Analysis

Below is the result from the mode file that svm-rank produce.

Exp2-2

```
1 1:0.66523343 2:-0.51012886 3:0.85268933 4:-0.039655864 5:0.41881642 6:0.043190911
7:0.23164007 8:0.39671803 9:0.0019704713 10:0.30618253 11:0.099441722 12:0.018023137
13:0.26893273 14:0.11295604 15:0.049165741 16:0.0065471395 #
```

Comb2:

```
1 5:1.0304322 6:0.26900321 17:1.2564716 18:-0.29500273 #
```

Comb3:

```
1 1:1.2964433 4:-0.066009223 5:1.0823898 17:1.0166085 #
```

Comb4:

```
1 1:1.0802242 5:0.74771911 10:0.71489888 17:1.0903645 #
```

Based on the above results, feature 1,3,5,7,8,13,17 (spam, Wikipedia, BM25 score for body, term overlap for body, BM25 for title, term overlap for url, document body length) has a relative good contribution to the accuracy. Theses features are more useful to support accuracy improvement. This can be a prove that in the combination 4, when include 1,5,10,17, all the variables have a relative good score. When look into these features, we can find that when having features with high accuracy, the features are not internal correlated correspondingly. The term overlap in url makes more sense than in body, and BM25 score always win when compare to Indri score, where Indri usually get negative or 0-closed score in the model. Also, one of my custom feature – document title length is not perform well in this case.