Homework 5

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: Diversity and relevance baselines

2.1 Experimental Results

	Indri (Exp-2.1a)	Indri + PM2 (Exp-2.1b)	Indri + xQuAD (Exp-2.1c)	BM25 (Exp-2.1d)	BM25+ PM2 (Exp-2.1e)	BM25+ xQuAD (Exp-2.1f)
P-IA@10	0.17166	0.21900	0.17416	0.23916	0.28050	0.29300
P-IA@20	0.20000	0.23491	0.20250	0.22541	0.30100	0.31650
αNDCG@20	0.31584	0.40616	0.31872	0.41438	0.49016	0.54762
P@10	0.2800	0.3400	0.2800	0.3900	0.4000	0.4200
P@20	0.3450	0.3700	0.3350	0.3850	0.4250	0.4550
MAP	0.2231	0.1382	0.1288	0.2666	0.1706	0.1828

2.2 Parameters

trecEvalOutputLength=100
retrievalAlgorithm=Indri
Indri:mu=1500
Indri:lambda=0.1
BM25:b=0.75
BM25:k_1=1.2
BM25:k_3=0.0

2.3 Discussion

Using diversity algorithm could lower the Accuracy score and MAP score. As we can see that MAP scores of 2.1b, 2.1c, 2.1e and 2.1f are lower than that of 2.1a and 2.1d. It is because the diversity algorithm will reorder the original ranking. The original ranking only optimizing for the original query and MAP, Precision are also calculated with respect to original query. Therefore, applying diversity algorithm would lower the score. But it is interesting to see that, although the MAP is lower when using diversity algorithm. P@10 and P@20 is not necessarily lower than the baseline. The reason could be that documents appearing on the top of the original ranking not only fit the original query well but also being able to fit all kinds of intents well. Therefore, these documents stay in the front when using diversity algorithm.

Using diversity algorithm improves P-IA and alpha-NDCG. And PM2 algorithm works better than xQuAD.

3 Experiment: Effect of λ

3.1 Experimental results

Indri + PM2				
	λ=0.0 (Exp-3.1a)	**λ=0.33 (Exp-3.1b)**	**λ=0.67 (Exp-3.1c)**	**λ=1.0 (Exp-3.1d)**
P-IA@10	0.21683	0.21933	0.21466	0.19550
P-IA@20	0.23841	0.24158	0.22825	0.20025
αNDCG@20	0.36887	0.41902	0.40410	0.40410
P@10	0.3000	0.3400	0.3400	0.3400
P@20	0.3500	0.3600	0.3600	0.3500
MAP	0.1407	0.1422	0.1342	0.1279
Indri + xQuAD				
	λ=0.0 (Exp-3.2a)	**λ=0.33 (Exp-3.2b)**	**λ=0.67 (Exp-3.2c)**	**λ=1.0 (Exp-3.2d)**
P-IA@10	0.17166	0.16666	0.18167	0.22183
P-IA@20	0.20000	0.20500	0.20875	0.23108
αNDCG@20	0.31584	0.31630	0.32084	0.40655
P@10	0.2800	0.2700	0.2900	0.3500
P@20	0.3450	0.3450	0.3350	0.3550
MAP	0.1213	0.1247	0.1323	0.1426

BM25 + PM2				
	λ=0.0 (Exp-3.3a)	**λ=0.33 (Exp-3.3b)**	**λ=0.67 (Exp-3.3c)**	**λ=1.0 (Exp-3.3d)**
P-IA@10	0.30533	0.27967	0.28633	0.19722
P-IA@20	0.30225	0.30725	0.29808	0.19514
αNDCG@20	0.48695	0.51528	0.53560	0.61049
P@10	0.4600	0.4100	0.4100	0.3000
P@20	0.4250	0.4350	0.4300	0.3500
MAP	0.1913	0.1747	0.1763	0.2800
BM25 + xQuAD				
	λ=0.0 (Exp-3.4a)	**λ=0.33 (Exp-3.4b)**	**λ=0.67 (Exp-3.4c)**	**λ=1.0 (Exp-3.4d)**
P-IA@10	0.23917	0.30050	0.29550	0.28133
P-IA@20	0.22542	0.29525	0.31650	0.31275
αNDCG@20	0.41438	0.55249	0.51235	0.49970
P@10	0.3900	0.4300	0.4300	0.4100
P@20	0.3850	0.4350	0.4550	0.4400
MAP	0.1787	0.1799	0.1796	0.1724

3.2 Parameters

Indri:mu=1500
Indri:lambda=0.1
BM25:b=0.75
BM25:k_1=1.2
BM25:k_3=0.0
diversity=true
diversity:algorithm=pm2
diversity:initialRankingFile=TEST_DIR/HW5-Exp-Indri.inRanks
diversity:maxInputRankingsLength=100
diversity:maxResultRankingLength=50

3.3 Discussion

For xQuAD, if lambda is higher, the algorithm will more focus on selecting diverse query. For PM2, if lambda is higher, the algorithm will tend to choose the document that fits a certain intent better, thus making the ranking more diverse. We assume that, if the lambda is higher, the ranking is more diverse and thus the MAP could be lower, because the ranking is not only optimized for matching the original query. Exp-3 seems to support this assumption, but other experiments do not. When we look at the P@10 or P@20, the score tends to be higher when lambda is higher. It means that when we consider more diversity in our ranking, the precision score gets higher. As for diversity metric, we need to tune the lambda between 0 and 1 to get the maximum score.

4 Experiment: The effect of the re-ranking depth

4.1 Experimental results

Indri + PM2				
	25 / 25 (Exp-4.1a)	**50 / 25 (Exp-4.1b)**	**100 / 50 (Exp-4.1c)**	**200 / 100 (Exp-4.1d)**
P-IA@10	0.20917	0.20500	0.21900	0.21967
P-IA@20	0.20750	0.21875	0.23492	0.25317
αNDCG@20	0.35386	0.35210	0.40616	0.46936
P@10	0.3300	0.3500	0.3400	0.3300
P@20	0.3400	0.3400	0.3700	0.3550
MAP	0.0724	0.0685	0.1382	0.2475

BM25 + PM2				
	25 / 25 (Exp-4.2a)	**50 / 25 (Exp-4.2b)**	**100 / 50 (Exp-4.2c)**	**200 / 100 (Exp-4.2d)**
P-IA@10	0.25583	0.30517	0.28050	0.27283
P-IA@20	0.23333	0.28258	0.30100	0.27992
αNDCG@20	0.48004	0.50670	0.49017	0.53431
P@10	0.3700	0.4500	0.4000	0.4100
P@20	0.3800	0.4250	0.4250	0.3900
MAP	0.1163	0.1185	0.1706	0.2812
BM25 + xQuAD				
	25 / 25 (Exp-4.3a)	**50 / 25 (Exp-4.3b)**	**100 / 50 (Exp-4.3c)**	**200 / 100 (Exp-4.3d)**
P-IA@10	0.26083	0.27917	0.29300	0.29700
P-IA@20	0.23833	0.29925	0.31650	0.29367
αNDCG@20	0.49658	0.49863	0.54763	0.51267
P@10	0.3900	0.4300	0.4200	0.4600
P@20	0.3800	0.4550	0.4550	0.4450
MAP	0.1508	0.1228	0.1828	0.3069

4.2 Parameters

Indri:mu=1500
Indri:lambda=0.1
BM25:b=0.75
BM25:k_1=1.2
BM25:k_3=0.0
diversity=true
diversity:algorithm=pm2 (or xQuad)
diversity:initialRankingFile=TEST_DIR/HW5-Exp-Indri.inRanks
diversity:lambda=0.5

4.3 Discussion

The deeper the re-ranking is, the higher score it gets, not only on MAP and precision scores but also on diversity metrics, P-IA and a-NDCG. If we use large re-ranking depth, we can have more candidates to form the final ranking, which can give us better result. But at the same time, it is possible that some bad documents also get included in the candidates, finally chosen into the result and damage the perform. We can see this on some of the P@10 and P@20 result. Overall, increasing the re-rank depth give us a better performance (within our range from 25 to 200), but the improvement is more obvious on diversity metric than the normal Precision and MAP scores.