Your Name Yiding Ou

Your Andrew ID yidingo

Homework 5

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

2. Did you give help of any kind to anyone in developing their software for this assignment (Yes or No)?
   No
   If you answered Yes, provide the name(s) of anyone that you helped, and describe the type of help that you provided.

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.
   Yes
   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.

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

Your Name Yiding Ou

Your Andrew ID yidingo

Homework 5

Instructions

1 Experiment: Diversity and relevance baselines

1.1 Experimental results

		Indri	Indri + PM2	Indri + xQuAD	BM25	BM25+ PM2	BM25+ xQuAD
Diversity	**P-IA@10**	0.080000	0.185833	0.090000	0.095833	0.298833	0.261333
	P-IA@20	0.124583	0.217500	0.143333	0.091667	0.258583	0.231083
	αNDCG@20	0.254028	0.392230	0.244209	0.253129	0.535741	0.512414
Relevance	**P@10**	0.1000	0.1500	0.1100	0.1200	0.3300	0.2900
	P@20	0.1550	0.2250	0.1650	0.1350	0.3050	0.2650
	P@30	0.1600	0.2467	0.2100	0.1433	0.2867	0.2767
	MAP	0.0318	0.0577	0.0434	0.0323	0.0984	0.0804

1.2 Parameters

For Indri, I use 0.4 for lambda, 2500 for mu

For BM25, I use 0.75 for b, 1.2 for k1, and 0 for k3.

For the diversity algorithm, I use 0.5 for lambda, 100 for max input ranking length, 50 for max result ranking length.

1.3 Discussion

From the overall score, when we use BM25 retrieval model, all score are relative higher than indri. I think the reason is the BM25 is a best-match retrieval model, so the retrieved documents of both intent and origin query have higher relevance. Since we have already considered all intents, best match retrieve model probably is a better choice for diversification system.

Either diversity algorithm retrieved a better result set than merely using the retrieval model. This means the intents are important to satisfy the information need, and origin query cannot explicitly shows the information need. This phenomenon especially shown on alph-NDCG, which means when we consider the intents, more relevance will retrieved and will satisfy user with different information need.

Between xquad and pm2 algorithm, the pm2 algorithm has better performance. The major difference between them is the pm2 algorithm does not consider the origin query. So I think the origin query is the reason of the difference in performance. Sometimes the query may show wrong information need, like the word with several meaning.

2 Experiment: Effect of λ

2.1 Experimental results

	λ=0.0	**λ=0.25**	**λ=0.50**	**λ=0.75**	**λ=1.0**
Indri + PM2					
P-IA@10	0.209167	0.199167	0.185833	0.175833	0.197833
P-IA@20	0.227500	0.230000	0.217500	0.212500	0.199333
αNDCG@20	0.424265	0.422309	0.392230	0.366196	0.379819
Indri + xQuAD					
P-IA@10	0.080000	0. 080000	0.090000	0.130000	0.190000
P-IA@20	0.124583	0.149167	0.143333	0.161250	0.224167
αNDCG@20	0.254028	0.238525	0.244209	0.255343	0.388986

	λ=0.0	**λ=0.25**	**λ=0.50**	**λ=0.75**	**λ=1.0**
BM25 + PM2					
P-IA@10	0.300500	0.301333	0.298833	0.294667	0.272667
P-IA@20	0.266333	0.256417	0.258583	0.257333	0.282500
αNDCG@20	0.537086	0.513472	0.535741	0.524557	0.514508
BM25 + xQuAD					
P-IA@10	0.095833	0.244667	0.261333	0.305500	0.303833
P-IA@20	0.091667	0.204833	0.231083	0.253583	0.259167
αNDCG@20	0.253129	0.496600	0.512414	0.519818	0.540242

2.2 Parameters

For Indri, I use 0.4 for lambda, 2500 for mu

For BM25, I use 0.75 for b, 1.2 for k1, and 0 for k3.

For the diversity algorithm, I use 0/0.25/0.5/0.75/1 for lambda, 100 for max input ranking length, 50 for max result ranking length.

2.3 Discussion

For PM2 with indri shows the best performance when lambda is equals to 0. For PM2 when the lambda is equals to 0, the intent covered will be 0. So this result probably shows for indri, the intent with highest quotient score is not the best choice, the quotient score is based on indri score, so the indri retrieval model is not suitable for queries with intents.

For PM2 with BM25 the score does not vary much between different lambda, when lambda is 0.5 the alpha-NDCG has highest score, this shows that lambda 0.5 is best for retrieval system that tend to cover all intents.

For xquad with both retrieval model shows better performance with larger lambda, the best result is retrieved when lambda is 1. When lambda is 1, the xquad will not consider the origin query. This also proved my previous argument, the original query sometimes has negative impact on retrieve result.

3 Experiment: The effect of the re-ranking depth

3.1 Experimental results

	25 / 25	50 / 25	100 / 50	200 / 100
		Indri + PM2		
P-IA@10	0.220000	0.201667	0.185833	0.199167
P-IA@20	0.238750	0.225917	0.217500	0.218750
αNDCG@20	0.415303	0.403921	0.392230	0.397572
		Indri + xQuAD		
P-IA@10	0.112500	0.105833	0.090000	0.090000
P-IA@20	0.154583	0.148333	0.143333	0.142917
αNDCG@20	0.247531	0.247584	0.244209	0.243739

	25 / 25	50 / 25	100 / 50	200 / 100
		BM25 + PM2		
P-IA@10	0.304167	0.305167	0.298833	0.333667
P-IA@20	0.251500	0.261750	0.258583	0.279083
αNDCG@20	0.566733	0.562652	0.535741	0.578297
		BM25 + xQuAD		
P-IA@10	0.205000	0.212667	0.261333	0.290333
P-IA@20	0.132917	0.173000	0.231083	0.239500
αNDCG@20	0.412612	0.473355	0.512414	0.551043

3.2 Parameters

The lambda I choosed is 0.5, since this value is the most neutral one for both algorithms and models. The bias will be minimized.

For Indri, I use 0.4 for lambda, 2500 for mu

For BM25, I use 0.75 for b, 1.2 for k1, and 0 for k3.

For the diversity algorithm, I use 0.5 for lambda, 25/50/100/200 for max input ranking length, 25/25/50/100 for max result ranking length.

3.3 Discussion

for indri with both PM2 and Xquad, the best result occurred when the di/dr is 25/25. This shows that the pm2 algorithm will has less documents to process, we will only give it the top 25 documents with scores that we retrieved with indri. And the more document we give to the pm2, the worse result we have. So this means that with indri model, top documents are more relevance to the information need, and the bottom documents are less relevance. And many relevance documents are not retrieved by indri. For the xquad, although the tendency is not clear, but the results still showed clues. The P-AI@10 and @20 dropped consistently.

For BM25 with both PM2 and xquad. The best result occurred when the di/dr is 200/100. Which means when we give more documents to the diversity algorithm, better result will be delivered. This is since BM25 is a best-match retrieve model, it has higher probability that the documents it retrieved document is a relevance document. Both P-AI@10 and @20 and alpha-NDCG@20 shows consistently increase with the di/dr increase.