Xiaoyu He

xiaoyuh1

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.

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

   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.
   Yes

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

Xiaoyu He

xiaoyuh1

Homework 5

1 Experiment: Diversity baselines

1.1 Experimental results

	Indri	Indri + PM2	Indri + xQuAD	BM25	BM25+ PM2	BM25+ xQuAD
P-IA@10	0.262167	0.303333	0.288667	0.183333	0.374333	0.383167
P-IA@20	0.268083	0.325667	0.305083	0.245000	0.333333	0.342833
αNDCG@20	0.462702	0.497056	0.474923	0.420487	0.607718	0.610105

1.2 Parameters

Indri: mu = 2500, lambda = 0.4
BM25: b = 0.75, k_1 = 1.2, k_1=0.0
diversity:maxInputRankingsLength = 100
diversity:maxResultRankingLength = 50
diversity:lambda = 0.5

1.3 Discussion

The results shows that diversification algorithm PM2 and xQuAD greatly improved the system's performance in both intent-aware precision and alpha-NDCG.

The improvement is significant for BM25, increased the P-IA@10 from 0.18 to 0.37, nearly doubled up. Compared to P-IA@10, base BM25 performs better in P-IA@20. Diversification increased P-IA@20 by 35%, alpha@NDCG by 45%.

For indri, the performance is not as significant as BM25. Indri-PM2 performs better than xQuAD. The PM2 increased P-IA@10 for Indri from 0.26 to 0.30, by 15%. P-IA@10 increased 14%, alpha-NDCG increase by 6%.

For BM25, PM2 and xQuAD have similar performance. For Indri, PM2 performs better than xQuAD. Why? The reason might due be the algorithm itself, but only one sample is not enough data draw this conclusion. Another more likely reason for different performance of PM2 and xQuAD on Indri and BM25 is the parameters we chosen for Indri and BM25. Changing the parameters may change the result.

2 Experiment: The effect of diversification on relevance

2.1 Experimental results

	Indri	Indri + PM2	Indri + xQuAD	BM25	BM25+ PM2	BM25+ xQuAD
P@10	0.3800	0.4500	0.4300	0.3300	0.5700	0.5800
P@20	0.4400	0.4700	0.4650	0.3750	0.4650	0.5050
P@30	0.4100	0.4300	0.4267	0.3667	0.4333	0.4500
MAP	0.2267	0.1968	0.1806	0.1997	0.2074	0.2191

2.2 Discussion

The following table summarized the result in a clearer way.

	Indri	Indri+PM2	Increased %	Indri +xQuAD	Increased %
P@10	0.38	0.45	16%	0.43	13%
P@20	0.44	0.47	6%	0.465	6%
P@30	0.41	0.43	5%	0.4267	4%
MAP	0.2267	0.1968	-15%	0.1806	-20%
	BM25	BM25+PM2	Increased %	BM25+xQuAD	Increased %
P@10	0.33	0.57	42%	0.58	76%
P@20	0.375	0.465	19%	0.505	35%
P@30	0.3667	0.4333	15%	0.45	23%
MAP	0.1997	0.2074	4%	0.2191	10%

For both indri and BM25, diversification greatly improved the precision.

For Indri, diversification improved P@10 by 16% but worsen the MAP score. This indicated that, diversification increased the precision but lowered the recall for Indri Model.

For BM25, diversification greatly improved P@10 by 40%-70%. Moving to the tail of ranking, from P@10 to P@30, the improvement in the precision gets smaller. Diversification increased the precision a lot by re-ranking the top documents. But re-rank in the “not so top” documents is not very helpful.

For BM25, xQuAD performs better than PM2 at all aspect. The reason is elaborated in the following paragraph. However, in Indri, PM2 and xQuAD have similar performance. Why? As I’ve already discussed in the previous section. The reason might due be the algorithm itself and the parameters we chosen for Indri and BM25.

3 Experiment: Effect of λ

3.1 Experimental results

	λ=0.0	λ=0.2	λ=0.4	λ=0.6	λ=0.8	λ=1.0
Indri + PM2						
P-IA@10	0.284000	0.289833	0.303333	0.310833	0.308833	0.290333
P-IA@20	0.315250	0.316917	0.328167	0.324583	0.318333	0.288583
αNDCG@20	0.506064	0.502698	0.496104	0.496729	0.499681	0.512333
Indri + xQuAD						
P-IA@10	0.262167	0.262167	0.272833	0.302000	0.289500	0.308500
P-IA@20	0.268083	0.289250	0.295917	0.305500	0.316417	0.324417
αNDCG@20	0.462702	0.487690	0.467385	0.471375	0.454257	0.498671

	λ=0.0	λ=0.2	λ=0.4	λ=0.6	λ=0.8	λ=1.0
BM25 + PM2						
P-IA@10	0.359167	0.365833	0.382000	0.369333	0.365167	0.303667
P-IA@20	0.327167	0.330167	0.331833	0.333167	0.338750	0.251833
αNDCG@20	0.596975	0.606256	0.598699	0.616105	0.620581	0.630049
BM25 + xQuAD						
P-IA@10	0.183333	0.358167	0.386667	0.388667	0.375333	0.367833
P-IA@20	0.245000	0.323917	0.333167	0.348000	0.326833	0.334583
αNDCG@20	0.420487	0.603042	0.610064	0.611670	0.605323	0.596128

3.2 Discussion

In order to discuss the result in a more clear way, here I pasted the result from baseline experiment 1.

	Indri	Indri + PM2	Indri + xQuAD	BM25	BM25+ PM2	BM25+ xQuAD
P-IA@10	0.262167	0.303333	0.288667	0.183333	0.374333	0.383167
P-IA@20	0.268083	0.325667	0.305083	0.245000	0.333333	0.342833
αNDCG@20	0.462702	0.497056	0.474923	0.420487	0.607718	0.610105

I also computed the min and max of scores when changing lambda (except lambda = 0).

	max	min	ratio
	Indri + PM2		
P-IA@10	0.310833	0.289833	7%
P-IA@20	0.328167	0.288583	14%
αNDCG@20	0.512333	0.496104	3%
	Indri + xQuAD		
P-IA@10	0.3085	0.262167	18%
P-IA@20	0.324417	0.28925	12%
αNDCG@20	0.498671	0.454257	10%

	BM25 + PM2		
P-IA@10	0.382	0.303667	26%
P-IA@20	0.33875	0.251833	35%
αNDCG@20	0.630049	0.598699	5%
	BM25 + xQuAD		
P-IA@10	0.388667	0.358167	9%
P-IA@20	0.348	0.323917	7%
αNDCG@20	0.61167	0.596128	3%

PM2:

Both indri and BM25 performs best when PM2 lambda=0.6. After choosing the desired intents, documents are re-ranked as following equation:

$$d^* = \arg\max_{d_j \in R}[\lambda qt[i^*]p(d_j|q_{i^*}) + (1-\lambda)\sum_{i \neq i^*} qt[i]p(d_j|q_i)]$$

The first part of this equation is the weight of covering current selected topic qi*, the second part is the weight of covering other intents. Lambda=0 means that, when ranking the documents, we only care if the document covers qi*, regardless of other intents.

For PM2, except lama = 0, alpha-NDCG is stable with regard to the chanage in lambda. The fluctuation for Indri is within 3%, BM25 within 5%.

The fluctuation in the P-IA@10 and P-IA@20 is relatively greater than alpha-NDCG. Especially for BM25. The best lambda=0.4, out performs the worst lambda=1 by 30% percent.

xQuAD:

The score of an document is calculated by the following equation:

$$d^* \leftarrow \arg\max_{d \in R \setminus S} (1-\lambda)\,\mathrm{P}(d|q) + \lambda\,\mathrm{P}(d, \bar{S}|q)$$

Lambda controls the weight of relevance score P(d|q) and diversity score P(d,S|d). lambda=1 means re-ranking only use the diversity score. lambda=0 means re-ranking documents only use the relevant score, this have the same effect of not using xQuAD at all. This is proved in our experiment result. As you can see from the table above, the blue shaded area has same results.

For xQuAD, except lama = 0, alpha-NDCG is stable with regard to the chanage in lambda. The fluctuation is within 3% for BM25, 10% for indri.

The fluctuation in the P-IA@10 is relatively greater, but not very much. For Indri-xQuAD, worst P-IA@10 is 0.262167, best is 0.302000, increase from worst to best is 15%. For BM25-xQuAD, worst P-IA@10 is 0.358167, best is 0.388667, increase from worst to best is 8%.

Indri: Best lambda for indri is 0.6 for the current parameter setting.

BM25: Best lambda for BM25 is 0.4-0.6 for the current parameter setting.

Conclusion:

The general trend is, huge fluctuation in P-IA@10 but stable reation in alpha-NDCG. This shows that lambda have more powerful impact on precision, on the top ranking documents.

Too big or too small lambda is not a good idea. Choosing a appropriate lambda is important and requires parameter tuning per parameter setting.

4 Experiment: The effect of the re-ranking depth

4.1 Parameters

According to above experiments, I chose lambda=0.6 for both xQuAD and PM2.

4.2 Experimental results

	25 / 25	50 / 25	100 / 25	100 / 50	200 / 100
Indri + PM2					
P-IA@10	0.318833	0.321667	0.310833	0.310833	0.300167
P-IA@20	0.272167	0.304417	0.324583	0.324583	0.323833
αNDCG@20	0.500071	0.477326	0.496729	0.496729	0.494035
Indri + xQuAD					
P-IA@10	0.277833	0.287167	0.302	0.302	0.280333
P-IA@20	0.280917	0.297583	0.3055	0.3055	0.30425
αNDCG@20	0.470559	0.468729	0.471375	0.471375	0.465182

	25 / 25	50 / 25	100 / 25	100 / 50	200 / 100
BM25 + PM2					
P-IA@10	0.3215	0.366	0.369333	0.369333	0.3735
P-IA@20	0.273167	0.3235	0.333167	0.333167	0.330833
αNDCG@20	0.61542	0.644895	0.616105	0.616105	0.604498
BM25 + xQuAD					
P-IA@10	0.325667	0.387667	0.388667	0.388667	0.383667
P-IA@20	0.268167	0.3275	0.348	0.348	0.344667
αNDCG@20	0.613842	0.676469	0.61167	0.61167	0.652761

4.3 Discussion

Stability:

The result is actually quit stable. The general trend is that, increase the number of input document increase the P-IA and alpha-NDCG. Although 25/25 is the worst, but it's not too bad. Still better than not using diversification.

Changing the number of input/output have more impact on P-IA@10 P-IA@20 compare to alpha-NDCG.

Using documents more than 50/25, the performance is actually really similar. Increasing the number of documents from 50 to 100 does not really improved the result much. Increase the input documents from 100 to 200 actually performs worse in Indri, but still, it doesn't change the performance a little bit.

Computation:

If we do not care about the computation, choosing 100/50 looks like the best option. But however, it's not the case in real life.

Considering the computation effort and the sweetness we get, choosing 50/25 looks like a most economic choice.