LIU Xi

xiliu1

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.
   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.

LIU Xi

xiliu1

Homework 4

1 Experiment 1: Baselines

Provide information about the effectiveness of your system in five baseline configurations.

	Ranked Boolean AND	Indri			
		BOW		Query Expansion	
		Your System	Reference System	Your System	Reference System
P@10	0.3250	0.2400	0.3100	0.2100	0.2350
P@20	0.3675	0.3075	0.3525	0.2800	0.3050
P@30	0.3700	0.3200	0.3467	0.2933	0.3250
MAP	0.1881	0.1760	0.1883	0.1659	0.1741
win/loss	N/A	12/8	14/6	11/9	13/7

Document the parameter settings that were used to obtain these results.

Indri BOW: mu=2500, lamda=0.4

Query Expansion: fbDocs=10, fbTerms=10, fbMu=0, fbOrigWeight=0.5

Comment on the quality and character of the query expansion terms that were included, and the weights that were produced. Do they seem reasonable?

Generally speaking, according to daily experiences, I could recognize in the expansions, some words appearing together with the query term very often or standing for similar meanings with the terms. However, some terms in the expansion seems less relevant. Overall, they seem reasonable.

As for weights, most of them are smaller than the original terms, and terms appearing together with the original terms more frequently were assigned larger weight, which seems reasonable.

Provide information about a few example queries to make your points, for example queries that had the most dramatic change in performance (good or bad) from query expansion, but do not provide information about every query individually. We are primarily interested in your observations about general trends, not quirky queries.

Overall, many expanded terms appears with query terms in daily life, or represent for similar meanings. For example, “djs” gets expansions like “entertainer”, “jockey”, “disc”, and “band”, which perfectly described its information needs, and helped significantly improved MAP to 0.2762.

For short queries, term in expansions sometimes look irrelevant. For example, “obama family tree” gets irrelevant expansions such as “crest” and “surname” with larger weight than relevant expansions such as

"genealogy" and "history", which probably a reason for expanded result (0.0169) losing to RankedBoolean result (0.0180).

For queries contains rare or professional terms, such as "uplift at yellowstone national park", it gets expansions such as "earthquake" and "volcano" which is related to "uplift", and "observatory" where described uplifts. That is probably why the expanded result (0.1234) gets higher MAP than RankedBoolean (0.0175).

Comment on the effects of query expansion on your system and on the reference system.

First of all, query expansion could improve MAP of some queries, but harm some queries at the same time.

Compared with the baseline, both expansions could introduce more improvements than harms. The expansion based on reference system performs better than my own Indri settings, in terms of win/loss ratio. However, compared with the algorithm they referring to, the two expansions are worse than their references.

Are the two systems affected equally by query expansion, or are there important differences?

For the sample queries, the two systems are both slightly affected by query expansion, that is MAP and top document precision is lowered.

As for differences, the terms and weights in expansions are different in the two systems. For example, "Obama family tree", both my system and the reference system chose genealogy (0.0744/0.0283), history (0.0553/0.0181), and surname (0.0761/0.0247), but their weights are different in two systems. Also, my system chose expansions such as engrave, gift, and trivia, which were not chosen by the reference system.

2 Experiment 2: The number of feedback documents

Provide information about the effect of the number of feedback documents on query expansion.

	Ranked Boolean AND	**Indri BOW, Your System**	**Query Expansion, Your Initial Results**					
			Feedback Documents					
			10	**20**	**30**	**40**	**50**	**100**
P@10	0.3250	0.2400	0.2100	0.2200	0.2150	0.2400	0.2350	0.2350
P@20	0.3675	0.3075	0.2800	0.2875	0.2975	0.2950	0.2950	0.2900
P@30	0.3700	0.3200	0.2933	0.3133	0.3100	0.3133	0.3217	0.3167
MAP	0.1881	0.1760	0.1659	0.1647	0.1674	0.1676	0.1688	0.1680
win/loss	N/A	12/8	11/9	13/7	11/8	12/8	12/8	10/9

	Ranked Boolean AND	Indri BOW, Reference System	Query Expansion, Reference System Initial Results					
			Feedback Documents					
			10	20	30	40	50	100
P@10	0.3250	0.3100	0.2350	0.2500	0.2150	0.2300	0.2400	0.2650
P@20	0.3675	0.3525	0.3050	0.3225	0.3100	0.3150	0.3150	0.3125
P@30	0.3700	0.3467	0.3250	0.3167	0.3200	0.3233	0.3200	0.3250
MAP	0.1881	0.1883	0.1741	0.1751	0.1718	0.1735	0.1743	0.1770
win/loss	N/A	14/6	13/7	13/7	12/8	12/8	12/8	12/8

Document the values of any parameters that were held constant during this experiment.

Indri BOW: mu=2500, lamda=0.4

Query Expansion: fbTerms=10, fbMu=0, fbOrigWeight=0.5

Comment on the effect of varying the number of feedback documents on the quality and character of the query expansion terms that were included, and the weights that were produced.

Different number of feedback documents could change the selection of expanding terms and weights for each term. As for the quality of expansion, more documents could help to find more related terms, and provide more evidence to evaluate weights. However, too many documents could harm the result, because some frequent and less relevant terms from lower-ranked documents could possibly be selected, also lower the weight of highly relevant terms.

Were any values consistently better than other values?

Yes. fbDoc = 50 outperforms others.

Does using more documents tend to help the results, or hurt the results? Why?

Yes. The general trend in both groups is that, the MAP and win/loss ratio increase when the number of feedback documents increases. Because more documents contain more related terms for expansion, and provide more evidences for estimation of the weights.

On the other hand, more documents introduces more candidates, which could possibly hurt the result because candidates appeared frequently in lower ranked documents may not related to query terms.

Provide information about a few example queries to make your points, for example queries that had the most dramatic change in performance as the number of documents varied, but do not provide information about every query individually. We are primarily interested in your observations about general trends, not quirky queries.

More documents could help to select more relevant terms. For example, in my own system, the expansion for "gmat prep" with fbDocs=50 outperforms the one with fbDocs=10. When fbDocs=10, it contains irrelevant terms such as "milic", "you" and "i"; while when fbDocs=50, it chose highly relevant terms such as "practice" and "gre", and higher weights are assigned to existing relevant terms such as course (0.0182/0.0599) and class (0.009/0.0305).

More documents could introduce some bad expansions. For example, in my own system, the expansion for "cheap internet" with fbDocs = 10 outperforms the RankedBoolean but lose to the it when fbDocs=50. When fbDocs=50, it selected irrelevant terms "levitra", "accutane", "viagra" and "ciali", which harms the performance.

If using more documents improves expansion quality, is the improvement worth the added computational costs?

No. The cost increased rapidly but MAP improves very slowly. For the expansions based on reference system, when fbDocs = 100, MAP only increased less than 0.003 compared with fbDocs = 10, which is not statistically significant, and the computation cost for sorting documents and storing terms increased more than 100 times.

Comment on the effects of query expansion on your system and on the reference system. Are the two systems affected equally by query expansion, or are there important differences?

For the sample queries, the two systems are both slightly affected by query expansion, that is MAP and top document precision is lowered.

As for differences, the terms and weights in expansions are slightly different in the two systems. For example. "Obama family tree", both my system and the reference system chose genealogy (0.0744/0.0283), history (0.0553/0.0181), and surname (0.0761/0.0247), but their weights are different in two systems. Also, my system chose expansions such as engrave, gift, and trivia, which were not chosen by the reference system.

3 Experiment 3: The number of feedback terms

Provide information about the effect of the number of feedback terms on query expansion.

	Ranked Boolean AND	Indri BOW, Your System	Query Expansion, Your Initial Results					
			Feedback Terms					
			5	10	20	30	40	50
P@10	0.3250	0.2400	0.2200	0.2100	0.2050	0.2050	0.2100	0.2150
P@20	0.3675	0.3075	0.2825	0.2800	0.2800	0.2850	0.2825	0.2825
P@30	0.3700	0.3200	0.2933	0.2933	0.3067	0.3017	0.3017	0.3050
MAP	0.1881	0.1760	0.1624	0.1659	0.1679	0.1688	0.1686	0.1689
Win/loss	N/A	12/8	9/11	11/9	12/8	11/9	11/9	12/8

	Ranked Boolean AND	Indri BOW, Reference System	Query Expansion, Reference System Initial Results					
			Feedback Terms					
			5	10	20	30	40	50
P@10	0.3250	0.3100	0.2150	0.2350	0.2300	0.2400	0.2450	0.2450
P@20	0.3675	0.3525	0.3025	0.3050	0.3025	0.3125	0.3025	0.3250
P@30	0.3700	0.3467	0.3067	0.3250	0.3133	0.3200	0.3200	0.3367
MAP	0.1881	0.1883	0.1667	0.1741	0.1640	0.1649	0.1651	0.1793
Win/loss	N/A	14/6	11/9	13/7	10/10	11/9	11/9	14/6

Document the values of any parameters that were held constant during this experiment.

Indri BOW: mu=2500, lamda=0.4

Query Expansion: fbDocs=10, fbMu=0, fbOrigWeight=0.5

Comment on the effect of varying the number of feedback terms on the quality and character of the query expansion terms that were included, and the weights that were produced.

Different number of feedback terms could change the selection of expanding terms and weights for each term.

As for the quality of expansion, more feedback terms could help to retrieve more relevant document, and provide more evidence to evaluate weights. However, too many terms could harm the result, because some less relevant terms were selected, and some highly relevant terms are under-weighted.

Were any values consistently better than other values?

Yes. fbTerms = 50.

Does using more terms tend to help the results, or hurt the results? Why?

Yes. The general trend in both groups is that, the MAP and win/loss ratio increase when the number of feedback terms increases. This is because more feedback terms help to retrieve more relevant document, and provide more evidence to evaluate weights.

On the other hand, more documents introduces more candidates, which could possibly hurt the result because candidates appeared frequently in lower ranked documents may not related to query terms.

Provide information about a few example queries to make your points, for example queries that had the most dramatic change in performance as the number of documents varied.

More feedback terms could help to retrieve more relevant document, and provide more evidence to evaluate weights. In the two systems, their MAP increases when the fbTerms increases.

Allowing more feedback terms could possibly expand the searching scope too large, and retrieve some irrelevant documents. For example, in my own system, the expansion for “mitchell college” with fbTerms = 5 (0.0736) outperforms the expansion with fbTerms=50 (0.0505). When fbTerms=50, it selected

irrelevant terms “depauw”, “foul”, “beauty” and “charles”, which harms the performance, and when fbTerms=5, only “depauw” was introduced to the expansion and assigned to very small weight (0.0037).

If using more terms improves expansion quality, is the improvement worth the added computational costs?

Yes. The cost is the same but MAP is improved. For the expansions based on reference system, when fbTerms = 50, MAP only increased by 13% compared with fbTerms = 5 and the computation cost for sorting terms are the same.

Comment on the effects of query expansion on your system and on the reference system. Are the two systems affected equally by query expansion, or are there important differences?

For the sample queries, the two systems are both slightly affected by query expansion, that is MAP and top document precision is lowered.

As for differences, the terms and weights in expansions are slightly different in the two systems. For example, for “mitchell college”, both my system and the reference system chose “university” (0.0059/ 0.0204), but their weights are different in two systems. Also, my system chose “biology”, which were not chosen by the reference system, instead, it chose “school”.

4 Experiment 4: Original query vs. expanded query

Provide information about the effect of varying the weight between the original query and the new expansion query.

	Ranked Boolean AND	Indri BOW, Your System	Query Expansion, Your Initial Results					
			fbOrigWeight					
			0.0	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
P@10	0.3250	0.2400	0.1900	0.2000	0.2050	0.2150	0.2350	0.2400
P@20	0.3675	0.3075	0.2825	0.2825	0.2800	0.2900	0.2875	0.3075
P@30	0.3700	0.3200	0.3000	0.3017	0.2983	0.2950	0.3150	0.3200
MAP	0.1881	0.1760	0.1490	0.1599	0.1653	0.1689	0.1752	0.1760
Win/loss	N/A	12/8	9/11	9/11	12/8	11/9	13/7	12/8

	Ranked Boolean AND	Indri BOW, Reference System	Query Expansion, Reference System Initial Results					
			fbOrigWeight					
			0.0	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
P@10	0.3250	0.3100	0.2250	0.2350	0.2400	0.2250	0.2250	0.2400
P@20	0.3675	0.3525	0.2975	0.3025	0.3125	0.3075	0.3000	0.3075
P@30	0.3700	0.3467	0.3150	0.3200	0.3183	0.3250	0.3283	0.3200
MAP	0.1881	0.1883	0.1636	0.1694	0.1718	0.1768	0.1776	0.1760
Win/loss	N/A	14/6	10/10	9/11	12/8	14/6	13/7	12/8

Document the values of any parameters that were held constant during this experiment.

Indri BOW: mu=2500, lamda=0.4

Query Expansion: fbDocs=10, fbTerms=10, fbMu=0

Comment also on the balance between the original query and the expansion query.

Generally speaking, in both systems, MAP increases when fbOrigWeight increases. This is because in terms of MAP, current expansions are worse than Indri Bow, so heavier weight on the Indri Bow score could make the final score more close to Indri-Bow score, that is, the retrieval results are more similar to Indri Bow results. If we take win/loss into consideration, fbOrigWeigh=0.8 is better.

Is a combination of the two queries worthwhile? Why or why not?

Yes. Expansions sometimes harm the result with irrelevant expansion terms or under-estimate weights of relevant expansion terms, so balancing the expansion with original query could help to reduce the harm.

How does the stability (win/loss) behavior compare to just using the expanded query alone?

In terms of win/loss ratio, for the expansions built on my system, combinations outperforms than expanded queries, except fbOrigWeight=0.2, which achieved the same win/loss ratio as the expansion did; for the expansions built on the reference system, combinations outperforms than expanded queries, except fbOrigWeight=0.2, which achieved a slightly lower win/loss ratio as the expansion did;

Comment on the effects of the combined query on your system and on the reference system. Are the two systems affected equally, or are there important differences?

For the sample queries, the two systems are both slightly affected by query expansion, that is MAP and top document precision is lowered. But for some queries, MAP is significantly improved, for example, among the combinations containing expansions built on my system, when fbOrigWeight=0.8, MAP of 13 out of 20 queries get improved.

As for differences, the terms and weights in expansions are slightly different in the two systems, but no important difference is observed.

5 Experiment 5: Effect of the original query quality

Provide information about the effect of varying the weight between the original query and the new expansion query

	Ranked Boolean AND	Query Expansion, Your Initial Results			
		BOW Original Query		SDM Original Query	
		Original	Expanded	Original	Expanded
P@10	0.3250	0.2400	0.2100	0.4200	0.3800
P@20	0.3675	0.3075	0.2800	0.4525	0.4125
P@30	0.3700	0.3200	0.2933	0.4567	0.4333
MAP	0.1881	0.1760	0.1659	0.2400	0.2404
Win/loss	N/A	12/8	11/9	15/5	16/4

	Ranked Boolean AND	Query Expansion, Reference System Initial Results			
		BOW Original Query		SDM Original Query	
		Original	Expanded	Original	Expanded
P@10	0.3250	0.3100	0.2350	0.4000	0.2600
P@20	0.3675	0.3525	0.3050	0.4200	0.3300
P@30	0.3700	0.3467	0.3250	0.4400	0.3417
MAP	0.1881	0.1883	0.1741	0.2231	0.1893
Win/loss	N/A	14/6	13/7	16/4	15/5

Document the values of the parameters used for this experiment.

Indri BOW: mu=2500, lamda=0.4

Indri SDM: w = 0.3

Query Expansion: fbDocs=10, fbTerms=10, fbMu=0, fbOrigWeight=0.5

Does a difference in the quality of the initial retrieval make any difference in query expansion effectiveness or stability?

Yes. SDM initial retrieval is of better quality than Indri Bow is, and in both expansions, expansions based SDM achieved better performance in terms of MAP and stability than expansions based on Indri Bow. This is probably because SDM is of better quality, in terms of MAP and top precision, using SDM as a reference could help find more relevant documents then extract more relevant documents and evaluate their weights more properly.

6 Analysis of results

You ran a lot of experiments, and have a lot of experimental results. The sections above discuss each experiment individually. In this section, we want you to think about general trends that you observed across the 5 experiments that have not been discussed in earlier sections.

How did query expansion affect the "high Precision" portion of a document ranking (the top-ranked documents) and the "high Recall" portion of the document ranking (farther down the ranking)?

For the high precision portion, noticing that p@10 p@20, and p@30 of expanded queries are actually lower than those of original queries in all experiments, we could conclude that, query expansion harms the precision for top-ranked documents.

For the high recall portion, noticing that MAP of expanded queries is actually lower than MAP of original queries in all experiments, we could conclude that, query expansion harms the precision for farther down the ranking.

This is because query expansion introduces more query terms, and if they are not highly related to the topic, it could lead to selecting irrelevant documents.

Where does query expansion have the greatest impact?

Query expansion brings greatest positive impact when the query is short but clearly expressed the information needs. For example, MAP of "solar panel" in RankedBoolean system is only 0.0336, and 0.2879 in my own Indri system, but after expansion (fbDocs=10, fbTerms=20), related terms such as "power", "system", "energy", "collector" are selected, and MAP increased to 0.3032. I think this is because in this type of cases, the reference is of relatively good quality because its information need is clearly represented, and based on good references, query expansion manages to select relevant terms to retrieve more relevant documents.

Was query expansion stable in your experiments (as indicated by the win/loss ratio)?

With same reference systems and same settings, the query expansion result is stable. For different settings, the win/loss ratio is stable with the scope of 11/9 – 13/7, except few cases.

Were any experimental conditions more or less stable?

The experimental conditions are stable because with the same settings, the reference systems listed exactly the same documents in the same order. For experiments 1 – 5, I used the same baseline and exactly the same setting for Indri, and same reference file for Indri Bow.

Was there a correlation between accuracy metrics and stability?

Yes, for most of cases, when the stability (win/loss ratio) is high, the accuracy is high too. This is because when we have many query result improved, MAP should be improved too. Besides, for some cases where the stability is higher but MAP is lower, we could observe that top precisions are actually improved.

Is the increased computational complexity worth the increased accuracy (if any)? Keep in mind that a "production" implementation of pseudo relevance feedback would be much more optimized and faster than your implementation.

For my implementation, due to the constraint of computational resources and limited references, the improvement brought by query expansion is very minor, while the time cost increases rapidly. For example, the system is over 10 times slower than the Bow system when fbDocs=10 and fbTerms=10, but MAP of expansions over the whole query set is actually lower than the MAP of original systems. Thus the increased computational complexity does not worth the increased accuracy.

For query expansions in the industry, which is more optimized and faster, I believe it worth the cost, since I have observe in my naïve implementation, MAP and top document precision of some queries are significantly improved by query expansion.

Feel free to include other comments about what you observed. You did a lot of experiments. This is your opportunity to let us know what you learned in this assignment.

In the experiment, I observed some queries clearly expressed the information need but achieve high MAP, and I believe it is probably because there are not sufficient documents in this corpus so that the system could not produce good original retrieval result, therefore could not select good terms for expansion. For example, for query "pampered chef" which stands for a local restaurant in Pittsburgh and few documents in the corpus contains the query terms, only gets MAP = 0.0043 with RankedBoolean, MAP = 0.0370 with Indri, and MAP = 0.0347 with fbDocs = 10 and fbTerms=10. Less relevant terms such as "magic", "personal", and "stone" were selected for expansion, which is probably the cause of drop in MAP and precision.