Implicit Feedback and Document Filtering for Retrieval Over Query Sessions

Ben Carterette, Praveen Chandar
University of Delaware
{carteret,pcr}@udel.edu

1. INTRODUCTION

The IR Lab at the University of Delaware participated in the 2011 Sessions track. The Sessions track features sequences of queries $q_1, ..., q_m$, with only q_m being the subject for automatic retrieval. There are four separate experimental conditions for q_m, each with a greater amount of data about user/system interaction for prior queries:

1. RL1: no interaction information; q_m only.
2. RL2: previous queries $q_1, ..., q_{m-1}$ known to the system.
3. RL3: previous queries and retrieved results known to the system.
4. RL4: previous queries, retrieved results, and clicks on retrieved results known to the system.

We used the different experimental conditions in the track to explore three research questions:

1. the effect of simple implicit feedback on retrieval results;
2. the effect of corpus filters on retrieval results;
3. the effect of duplicate detection and removal on retrieval results.

2. METHODS

We used the same Indri index of ClueWeb09 that we built and used for last year's TREC submissions [1].

All of our queries use the Indri query language. When we did not use feedback, we used a simple keyword query, resulting in scoring by a Dirichlet-smoothed language model. When we did use feedback, we used a weighted combination of the original query and weighted expansion terms derived from the feedback documents. An example is: #weight(0.8 #combine(peace corp application) 0.2 #weight(0.055 corps 0.054 peace 0.051 volunteer 0.037 peacecorp 0.035 kennedy 0.031 benefit 0.030 application 0.029 president 0.028 info))

2.1 Implicit feedback

The RL4 condition provided click data for results retrieved prior to the current query. We used this data as implicit feedback about relevance with which to expand the original query:

1. expansion based on clicked documents only;
2. expansion based on unclicked documents only;
3. expansion based on all retrieved documents.

We expanded in a relatively simple way, calculating tf-idf weights for terms in the documents chosen for feedback, then using those tf-idf weights as term weights in an Indri query. The expanded query gave 4 times as much weight to the original query as to the expansion terms; this is based on decent results from previous experiments.

2.2 Corpus filtering

We used three different filters for the ClueWeb09 collection: a spam filter for the full collection based on the University of Waterloo spam scores [2], a "category B" filter that limits retrieved results to the first English-language subdirectory, and a Wikipedia filter that limits retrieved results to Wikipedia pages.

The latter two of these only involved querying a subset of the index, and therefore do not need to be described further.

We filtered spam as a post-retrieval step: all pages were indexed, but after retrieving documents we removed any that had a Waterloo spam score percentile of 0.75 or lower.

2.3 Duplicate document filtering

Session track sessions include a sequence of queries $q_1, ..., q_{m-1}$ leading up to the current query q_m that is the subject of experimentation. It is possible (and likely) that results retrieved for a query q_i will be retrieved again for a later query. Whether a user would want to see such results is an open question (and probably depends a great deal on context). We wanted to see how effectiveness would be affected if duplicates were removed.

We used two different methods for this depending on the available data:

- For the RL3 condition, we are given the previous queries as well as retrieved results (by a custom search engine built for the Session track). We filtered from the results for q_m any document that appeared in any ranking for the previous queries.
- For the RL2 condition, we are only given the previous queries with no retrieved results. We submitted these queries to our own Indri system, and then filtered from the results for q_m any document that appeared in the top 10 ranked results for the previous queries.

We expect the latter to result in more documents being filtered than the former.

.

feedback	nDCG@10	ERR	AP	GAP
no feedback	0.3201	0.2581	0.1346	0.1279
all docs	0.3904	0.3040	0.1583	0.1527
all clicked	0.3871	0.3040	0.1575	0.1528
all unclicked	0.3904	0.3082	0.1575	0.1518

Table 2: nDCG@10, ERR, AP, and GAP results for different implicit feedback methods. Feedback works, but the way the feedback documents are chosen doesn't seem to matter.

corpus	baseline	indri filter	track filter
cat A/spam	0.3201	0.3148	0.3180
cat B	0.2769	0.2904	0.2722
Wikipedia	0.3743	0.2767	0.3675

Table 3: nDCG@10 results for different corpus and duplicate document filters. Retrieving only Wikipedia pages is a good approach, but less so if anything will be filtered.

3. EXPERIMENTS

We submitted three runs consisting of four input files each (for each of the four experimental conditions RL1, RL2, RL3, and RL4). Details are given in Table 1.

We use these runs to test the experiments listed above as follows:

- **experiment 1:** compare implicit feedback with clicked documents and implicit feedback with unclicked documents to baselines of no feedback and feedback with all documents.

  The runs for this experiment are udelASFe1new.RL1, udelASFe1new.RL4, udelBe2.RL4, and udelWPmnz.RL4.

- **experiment 2:** compare three "corpus filters" on the full ClueWeb09: a spam filter on the full English-language collection, a "category B" filter limiting retrieved results to only the first subdirectory of English-language pages, and a Wikipedia filter further limiting retrieved results to only the Wikipedia crawl included in the collection.

  The runs for this experiment are udelASFe1new.RL1, udelBe2.RL1, and udelWPmnz.RL1.

- **experiment 3:** compare two different duplicate filtering methods: one that filters duplicate results that would have been retrieved by our own system for the previous queries in the session, and one that filters duplicate results retrieved by the Session track system.

  The runs for this experiment are udelASFe1new.RL1, udelASFe1new.RL2, udelASFe1new.RL3; udelBe2.RL1, udelBe2.RL2, udelBe2.RL3; and udelWPmnz.RL1, udelWPmnz.RL2, and udelWPmnz.RL3.

3.1 Results

3.1.1 Implicit feedback

Results for implicit feedback experiments are shown in Table 2. Interestingly, every feedback approach gave a big gain over the RL1 baseline for every measure we looked at. The obvious question is whether this gain is because we were expanding the query using results from a "good" search engine (the one that retrieved the documents in the RL3 condition); to test this, we should compare to feedback with documents retrieved by our Indri system. We have yet to do this.

There is no clear difference between the three feedback methods, though. Only small differences can be observed. It is interesting that the rank ordering of the results varies depending on evaluation measure: for nDCG@10, using clicked documents hurts performance, while for graded AP it provides a slight gain.

Our conclusion here is that feedback based on previous ranked results works well (though there may be a confounding effect), but which documents are used for feedback don't matter.

3.1.2 Corpus and duplicate filtering

It is immediately clear from looking at the first column of Table 3 that retrieving only Wikipedia documents provides the best overall effectiveness. This is likely because Session track topics, which were informational in nature, are well-matched to Wikipedia.

The effectiveness of retrieving from category A with spam filtering is still fairly good, and could likely be improved by giving more weight to Wikipedia pages. Retrieval in category B only is clearly the worst, though still fairly good considering it is a small subset of the entire collection.

Results from applying duplicate filtering depend heavily on both corpus filter and duplicate filtering method. First, regardless of corpus filter, using the filter based on RL3 data gives very little change in effectiveness. The simplest explanation for this is that few documents are actually being filtered; this would be the case if the RL3 results and our Indri system tend to retrieve very different results. Indeed, the mean overlap between the two systems (defined as the size of the intersection of their retrieved results divided by the size of the union) is only 0.03 even in the most direct comparison (spam-filtered category A); they retrieved almost no documents in common.

The Indri-based filter for the RL2 condition, on the other hand, sometimes makes a big difference and sometimes very little difference at all. In the Wikipedia case, filtering against Indri results in a very large decrease in effectiveness. This is likely because there is a fairly small subset of Wikipedia pages that are relevant and being retrieved for query after query; filtering them has a strong negative effect on effectiveness. This does raise the question: would a user really want to see the same Wikipedia page more than once in a session?

In the category B case, the filter has a small but positive effect. This may be because effectiveness in category B is relatively low, so nonrelevant documents tend to get filtered more than relevant documents.

In category A, the filter has a small negative effect. Contra the category B case, this may be because effectiveness in category B is higher, so more relevant documents are being filtered.

4. CONCLUSION

We used the Session track to explore three questions regarding implicit feedback and filtering. Results at this point are somewhat inconclusive. But it seems very clear that we must investigate feedback further, in particular the differ-

run name	condition	description
udelASFe1new	RL1	ad hoc retrieval on spam-filtered category A (ASF) index
	RL2	RL1 + removal of duplicates as identified by indri system
	RL3	RL1 + removal of duplicates as identified by track data
	RL4	implicit feedback by expanding query with clicked documents from previous interaction
udelBe2	RL1	ad hoc retrieval on category B index
	RL2	RL1 + removal of duplicates as identified by indri system
	RL3	RL1 + removal of duplicates as identified by track data
	RL4	implicit feedback for ASF by expanding query with unclicked documents from previous interaction
udelWPmnz	RL1	ad hoc retrieval on Wikipedia (enwp*) index
	RL2	RL1 + removal of duplicates as identified by indri system
	RL3	RL1 + removal of duplicates as identified by track data
	RL4	implicit feedback for ASF by expanding query with all documents from previous interaction

Table 1: Description of submitted runs. Note that udelBe2.RL4 and udelWPmnz.RL4 are part of an experiment involving the spam-filtered category A index, not the category B or Wikipedia indexes that their names suggest.

ence in effectiveness between feedback with our Indri system and feedback with the custom system used for the Session track. Since these two systems have very little overlap in retrieved results, it is possible that all gain in Table 2 is due to the former system being of high quality. This would call into question the validity of the test collection formed for the Session track.

The Session track also provided subtopic judgments that we have not had a chance to investigate yet. It is likely that we could use these for diversity across a session, however, and this is definitely a direction we intend to pursue.

5. REFERENCES

[1] Ben Carterette and Praveen Chandar. Sessions, diversity, and ad hoc retrieval. In *Proceedings of TREC*, 2010.

[2] Gordon V. Cormack, Mark D. Smucker, and Charles L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *CoRR*, abs/1004.5168, 2010.