U Albany & USC at the TREC 2012 Session Track

Xiaojun Yuan[1], Jingjing Liu[2], Ning Sa[1]

[1] College of Computing and Information, University at Albany, The State University of New York, Albany, NY 12222, (xyuan, nsa)@albany.edu
[2]School of Library and Information Science, University of South Carolina, Columbia, SC 29208, jingjing@mailbox.sc.edu

1. Introduction

Research in Information Retrieval has noticed that it is often the case that in a search process, users begin an interaction with a sufficiently under-specified query and they will need to reformulate their query multiple times before they find desired information. In terms of this, researchers hypothesized that a search engine may be able to better serve a user "by ranking results that help "point the way" to what the user is really looking for, or by complementing results from previous queries in the sequence with new results, or in other currently-unanticipated ways." (Kanoulas, Carterettey, Hallz, Cloughx, & Sanderson, 2011). In 2012, the goal of session track is: (G1) to test whether system performance can be improved for a given query by using previous queries and user interactions with the system (including clicks on ranked results, dwell times, etc.), and (G2) to evaluate system performance over an entire query session instead of a single query. Our UAlbany and USC group joined the Session Track task by taking into account searcher behaviors during the course of their information seeking process. In the following, we give an overview of our approach, details of our submission runs, our results, and our conclusions about the results.

2. Method: the prediction models, queries, and runs

2.1. RL1-Baseline

For the baseline, we used pseudo relevance feedback built on the current query, i.e., the last query in a session. We used the default parameters in Indri. This method was used by TREC 2011 session track participants such as Rutgers University (Liu et al., 2011).

The following describes the specific parameters:

```
<fbDocs>10</fbDocs>
<fbTerms>10</fbTerms>
<fbOrigWeight>0.5</fbOrigWeight>
<fbMu>0</fbMu>
```

After generating the queries, the symbols like "." in an URL and "'" in "Newton's law" were manually converted to 64 based string because the search engine did not recognize them.

2.2. RL2-queries

Our RL2 considered all queries in the current session. We used the findings in Liu et al. (2010) about the performance of query reformulation type to determine if a reformulated query was good or not. Liu et al. (2010) had different query formulation types, as described in Table 1. It was found that Generalization did not lead to useful pages, therefore, in our current approach, those reformulation type being Generalization was treated as "bad" queries, and all others being "good" queries.

Table 1. Query reformulation types (after Liu et al. (2010))

Type	Definition	Example
Generalization	Qi and Qi+1 contain at least one term in common; Qi+1 contains fewer terms than Qi	Qi: russian submarine Kursk on-board commander Qi+1: kursk commander Session 11
Specialization	Qi and Qi+1 contain at least one term in common; Qi+1 contains more terms than Qi	Qi: pocono mountains Qi+1: pocono mountains park Session 5
Word substitution	Qi and Qi+1 contain at least one term in common; Qi+1 has the same length as Qi, but contains some terms that are not in Qi.	Qi: pocono mountains camelbeach hotel Qi+1: pocono mountains chateau resort Session 6
Repeat	Qi and Qi+1 contain exactly the same terms, but the format of these terms may be different	Qi: Kursk Barents see Russian politics causes Qi+1: Kursk Barents see Russian politics causes Session 16
New	Qi and Qi+1 do not contain any common terms	Qi: pocono resort Qi+1: skytop lodge directions Session 7

Note: Qi+1: the query immediately following Qi in the current session

In our approach, the adjacent two queries were compared and good queries were picked. The first query was always taken as good. All the good queries were used to expand the current query. Again, some symbols were converted to 64 based format manually and pseudo relevant feedback was used.

2.3. RL3

The RL3 run considered all queries issued and webpages viewed in the current session. Determining if a query is "good" or not followed the same way as was done in the RL2 run. For the viewed webpages, our approach was to treat those whose titles match the query non-stop terms as "good" pages; otherwise, bad pages. We also treated Wikipedia pages as good ones.

In the run, firstly good queries were picked based on the same criteria used in RL2. Then the results of the good queries were examined. If the title of the result page matched the query non-stop terms, the page was taken as good page. Wikipedia pages were taken as good pages. The titles of the good pages and the queries of the good queries were used to expand the current query. Some symbols were converted to 64 based format manually and pseudo relevant feedback was used.

2.4. RL4

The RL4 run considered queries issued, webpages viewed, as well the users' interaction behaviors. The query and webpage approaches were the same as described above in the RL3 run. For other variables and criteria used, we referred to Agichtein (2006).

Specifically, good queries were picked. Then clicked results of the good queries were examined. If the title, URL and snippet of the clicked result all matched the query non-stop terms, the result page was considered as a good page candidate. The time spent on the adjacent three good page candidates was compared and the page with the longest dwell time was picked as a good page. If a good query did not have clicked result, the method in RL2 was used to pick good pages. The current query was expanded accordingly. Some symbols were manually converted to 64 based and pseudo relevant feedback was used.

3. Results

3.1. Mean of our results

Table 1 and Figure 1 show the means of various evaluation measures of the 96 tasks for all our 4 runs. Table 2 and Figure 2 show the means of various measures of the 48 topics for all our 4 runs. In general, as can be seen, across various measures, RL2 was roughly the same as RL1, RL3 was the best out of all our 4 runs, and RL4 was slightly better than RL1 and RL2, but not as well as RL3.

We found that the measures for 98 tasks had very similar patterns with the 48 topics. So in the following analysis, we used the 98 tasks only.

Table 1. Mean measures of 4 runs of 98 tasks

Measure	RL1	RL2	RL3	RL4
Average precision	0.1153	0.1151	0.1254	0.107
err	**0.1016**	**0.1019**	**0.1275**	**0.1085**
ERR@10	0.0904	0.0889	0.1191	0.0989
nDCG	0.2796	0.2893	0.2655	0.2603
nDCG@10	**0.1407**	**0.1294**	**0.1763**	**0.1409**
nerr	0.173	0.1756	0.2136	0.1819
nERR@10	0.1529	0.1524	0.1987	0.1652
Precision@10	**0.2296**	**0.2327**	**0.3357**	**0.2398**

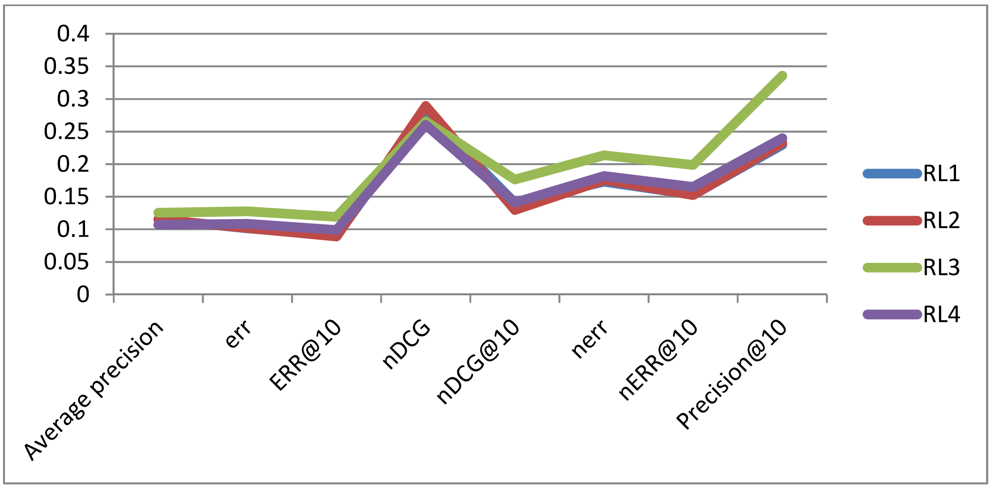

Figure 1. Mean measures of 4 runs of 98 tasks

Table 2. Mean measures of 4 runs of 48 topics

Measure	**RL1**	**RL2**	**RL3**	**RL4**
Average precision	0.1156	0.1158	0.1219	0.0988
err	**0.1035**	**0.1054**	**0.1217**	**0.1038**
ERR@10	0.0923	0.0920	0.1130	0.0939
nDCG	0.2884	0.3011	0.2618	0.2625
nDCG@10	**0.1439**	**0.1300**	**0.1649**	**0.1351**
nerr	0.1764	0.1807	0.2046	0.1768
nERR@10	0.1568	0.1570	0.1890	0.1601
Precision@10	**0.2365**	**0.2458**	**0.3438**	**0.2449**

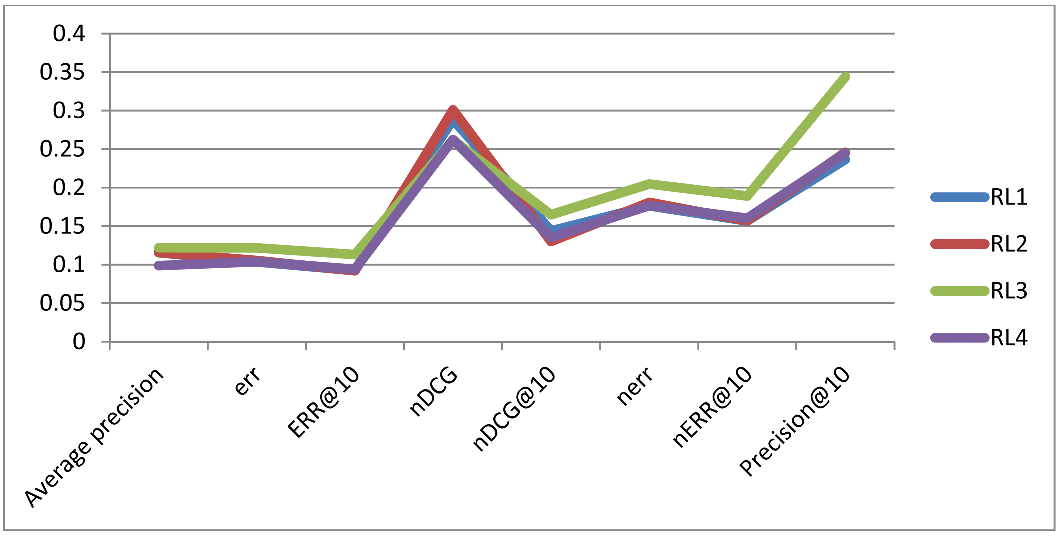

Figure 2. Mean measures of 4 runs of 48 topics

3.2. Changes over baseline run RL1

We further examined the changes (improvement/decrease) of our models over our baseline RL1. We used ERR (Expected Reciprocal Rank), ndcg@10, and precision@10 for evaluation of our results. ERR was selected because it was based on the "cascade" user model, ndcg@10 because it is the "basic" evaluation measure for the track, and precision@10 was used because it was a frequently used measure with direct evaluation of how the system does in its 1st SERP. Table 3 shows the results.

As can be seen, RL3 and RL4 outperformed our baseline in all measures. RL2 improved in Err and precision@10 but not nDCG@10.

Table 3. Changes over RL1

	ERR		nDCG@10		precision@10	
	Absolute improvement	Percent improvement	Absolute improvement	Percent improvement	Absolute improvement	Percent improvement
RL2	0.0003	0.2952	-0.0113	-8.031	0.0031	1.350
RL3	0.0259	25.49	0.0356	25.30	0.1061	46.21
RL4	0.0069	6.791	0.0002	0.1421	0.0102	4.442

3.3. Comparison between RL2, RL3, and RL4

We also compared the performance of all our models, RL2, RL3, and RL4. Again, we used err, ndcg@10, and precision@10 in this comparison. Table 4 shows the results.

As can be seen, RL3 improved much than RL2, with 25.1% in Err measure, 36.2% in nDCG@10, and 44.2% in precision@10. RL4 also improved over RL2, and the ratio was 6.5% in Err, 8.9 in nDCG@10, and 3.0% in precision@10. A direct comparison of RL4 over RL3 received unsurprising decrease in all 3 measures, dropping 14.9% in Err, 20.1% in nDCG@10, and 28.6% in precision@10.

Table 4. Comparison between RL2, RL3, and RL4

		Changes over RL2		**Changes over RL3**	
		Absolute improvement	Percent improvement	Absolute improvement	Percent improvement
Err	**RL3**	**0.0256**	**25.12**		
	RL4	**0.0066**	**6.48**	**-0.019**	**-14.9**
nDCG@10	**RL3**	**0.0469**	**36.24**		
	RL4	**0.0115**	**8.89**	**-0.0354**	**-20.08**
precision@10	**RL3**	**0.103**	**44.26**		
	RL4	**0.0071**	**3.05**	**-0.0959**	**-28.56**

4. Discussion/Conclusions

In our baseline, based on standard Indri techniques, we used pseudo relevance feedback built on the current query, i.e., the last query in a session. For the experimental runs, we used the findings in Liu et al. (2010) about the performance of query reformulation type to determine good queries in RL2, and then used the comparison between titles of viewed webpages and their queries to determine the good webpages in RL3. In RL4, we adopted the model proposed by Agichtein (2006) to fully take into account issued queries, viewed webpages and the users' interaction behaviors. We discovered that, in general, and evaluated by ERR, ndcg@10, and precision@10, RL3 run performed best than the other three runs. Our RL4 run was better than RL2 run in terms of err, and precision@10, but not for the rest of the measures. The results may be attributed to the various task features and types since the RL3 run identified a good webpage if there was a match between the title and the query.

In the future, we plan to further explore the relationship between the task types, task difficulties, domain types, and users' search behavior, as well as the appropriate or better ways of computing the usefulness scores for each webpage.

5. Acknowledgments

This research is sponsored by the Institute of Museum and Library Services grant (RE-04-10-0053-10).

References

Agichtein, E. & Zheng, Z. (2006). Identifying "best bet" web search results by mining past user behavior. *Proceedings of KDD '06*.

Kanoulas, E., Carterettey, B., Hallz, M., Cloughx, P., Sanderson, M. (2011). Overview of the TREC 2011 Session Track.

Liu, C., Gwizdka, J., Liu, J., Xu, T., and Belkin, N.J. (2010). Analysis and evaluation of query reformulations in different task types. *Proceedings of ASIS&T '10*.

Liu, C., Sun, S., Cole, M., & Belkin, N. J. (2011). Rutgers at the TREC 2011 Session Track. TREC 2011.