PITT at TREC 2011 Session Track

Jiepu Jiang, Shuguang Han, Jia Wu, Daqing He
School of Information Sciences,
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15260, USA
{jij29, shh69, jiw66, dah44}@pitt.edu

ABSTRACT
In this paper, we introduce our approaches for TREC 2011 session track. Our approaches focus on combining different query language models to model information needs in a search session. In RL1 stage, we build ad hoc retrieval system using sequential dependence model (SDM) on current query. In RL2 stage, we build query language models by combining SDM features (e.g. single term, ordered phrase, and unordered phrase) in both current query and previous queries in the session, which can significantly improve search performance. In RL3 and RL4, we combine query model in RL2 with two different pseudo-relevance feedback query models: in RL3, we use top ranked Wikipedia documents from RL2's results as pseudo-relevant documents; in RL4, snippets of the documents clicked by users in a search session are used. Our evaluation results indicate: texts of previous queries in a session are effective resources for estimating query models and improving search performance; mixing query model in RL2 with the query model estimated using click-through data (in RL4) can improve performance in evaluation setting that considers all subtopics, but no improvement is observed in evaluation setting that considers the only subtopic of current query; our methods of mixing query model in RL2 with query model in RL3 did not improve search performance over RL2 in any of the two evaluation settings.

Keywords
TREC, session, query language model, relevance feedback.

1. INTRODUCTION
In a recently study, Trotman and Keeler [1] find that search engines can provide ranked results comparable to those of human assessors (who are not those produced official judgments) on several TREC and INEX collections. They claim that there may be "not much room for improvement" for ad hoc information retrieval. Besides, it is also indicated from their results (in Fig. 2 of [1]) that the judgments from human assessors are far from perfect (with the medium of average precision ranging from about 0.2 to 0.6 on different collections). This demonstrates that context information associated with a search query is important, without it, it is even difficult for human beings to correctly understand the underlying information needs of a query (from another person).

From this point of view, TREC session track provides a platform with open collections and evaluation environments for studying retrieval techniques with context information - different layers of historical user behaviors in a search session. In 2011 session track tasks, four layers of session information are provided:

RL1: only current query (ad hoc retrieval setting)

RL2: current query and past queries in the same session

RL3: RL2 and top returned documents for each of the past queries

RL4: RL3 and user's click-through for each of the past queries

For each search topic, a few subtopics are identified and assigned to queries of the topic (which cannot be used for search). Because there may be transition of subtopics among queries in a search session, two different evaluation approaches are adopted in 2011 session track: one evaluates by documents relevant to any of the subtopics or the general topic ("all-subtopics" evaluation); the other evaluates by only documents relevant to the subtopic of current query ("current-only" evaluation). For details of session track task settings and evaluation methods, please refer to [2] and session track overview of the year.

Our retrieval approaches focus on using different mixture of query language models to incorporate into our retrieval process various session context information. In RL1 stage, we build ad hoc search system using sequential dependence model on current query. We further combine current query with two kinds of session context information and create correspondent query models for retrieval. One kind of session context information is the texts of users' past queries: we combine sequential dependence model query features (including single term, ordered phrase, and unordered phrase) in current query with those in past queries in RL2 stage. The other kind of session context is pseudo-relevance feedback information in the session, for which two different sets of documents are used as pseudo-relevance feedback documents set: in RL3, we use top ranked Wikipedia documents from RL2's results; in RL4, we use snippets of the documents clicked by users. The pseudo-relevance feedback models are combined with RL2's query model in search.

The rest of the article is organized as follows: section 2 introduces the methods we used for session track and some necessary details of implementation; in section 3, we analyze evaluation results; we finally discuss future works and draw a conclusion in section 4.

2. METHODS
2.1 System and Collection
We use Indri 5.0 to build our retrieval system because lots of our methods can be quickly implemented using Indri query language. We use only Clueweb09 category B collection [3] in experiments.

2.2 Ad hoc Retrieval (RL1)
In RL1 stage, we use sequential dependence model (SDM) [4] for ad hoc retrieval. We extract SDM features (including single term, ordered phrase, and unordered phrase) from current query and use Indri query language to incorporate SDM features into retrieval.

Besides, we also noticed from previous studies in web track that Clueweb09 collection involves lots of spam documents. Thus, we use Waterloo spam ranking score [5] to filter spam documents. In all our submitted runs, we first return top 5,000 documents and then filter out those with "fusion" spam score [6] less than 70%. This way of implementing filtering, however, does not guarantee to return 2,000 documents (the maximum number for evaluation)

for each topic, which may lead to lower MAP evaluation results than it should be in all our runs (but nDCG@10 will be accurate).

We tune SDM feature weights and spam score threshold on 2010 session track RL2 data (in 2010, RL2 uses only current query) to maximize nDCG@10. Table 1 shows results of query likelihood and SDM on 2010 session track data.

2.3 Session Historical Query Model (RL2)

In RL2 stage, except for current query, previous search queries in the same session can be used for ranking (which corresponds to the setting of RL3 in 2010 session track). We noticed a simple but effective method in 2010 session track for this setting: in [7] and [8], current query and a previous query were combined as a new query for search, which improved nDCG@10 in both studies. We also employ the idea in our RL2 run and expand it to incorporate SDM features.

In language modeling approaches for information retrieval, query model was firstly introduced in risk minimization framework [9] to model user's information needs, which is defined as generative model for queries. In an ad hoc retrieval setting, we can only have current query, a very limited sample from the query model. Thus, alternative samples were usually adopted for estimation of query models, e.g. pseudo-relevant documents [10, 11]. Now in session track settings, we can have previous search queries as alternative samples for query model estimation (although the sample size is still small). However, it is common that user's information needs may evolve during a search session. As a result, we may need to discount past queries in query model estimation.

Our method for RL2 stage (will be referred to as session historical query model (SH-QM) in following texts) estimates query model based on texts of current query and historical queries in the search session. Let F be a specific type of query feature, in our method, single term (T), ordered phrase (O), or unordered phrase (U). For any type of feature F, the probability of generating f, a specific feature value, is estimated as (1): q_{m} stands for current query and q_1 to $q_{\text{m-1}}$ stand for previous queries; $p_{MLE}(f|q_m)$ is the maximum likelihood estimation of f from q_{m}, which is calculated as (2); $p_{MLE}(f|q_1, \dots, q_{m-1})$ is the maximum likelihood estimation of f from past queries, as in (3); λ is the weight for previous queries. In (2) and (3), $count(f, q)$ is the raw frequency of feature value f observed in query q. Finally, three types of features are combined (using the tuned feature weights from 2010 data) for search.

$$\hat{p}_{SH}(f|q) = (1-\lambda) \cdot p_{MLE}(f|q_m) + \lambda \cdot p_{MLE}(f|q_1, \dots, q_{m-1}) \quad (1)$$

$$p_{MLE}(f|q_m) = \frac{count(f, q_m)}{\sum_{f'} count(f', q_m)} \quad (2)$$

$$p_{MLE}(f|q_1, \dots, q_{m-1}) = \frac{\sum_{i=1}^{m-1} count(f, q_i)}{\sum_{f'} \sum_{i=1}^{m-1} count(f', q_i)} \quad (3)$$

When considering only feature T (single term) and setting λ to 0.5, (1) is similar to the approaches adopted by [7] and [8]. We tune λ using RL3 data of 2010 session track (in 2010, RL3 uses both q_{m} and $q_{\text{m-1}}$ for retrieval) to maximize nDCG@10. The tuned weight for λ is 0.3, which also indicates we should discount past queries. Table 1 shows nDCG@10 results of SH-QM in 2010 session track data and comparison with results using only current query. It is indicated from the results that SH-QM can largely improve the performance compared with using only current query for search, no matter only single term feature or all SDM features are used. Besides, table 1 also indicates feature O and U can further slightly improve the performance over methods using only T in SH-QM.

Table 1. nDCG@10 of SH-QM using 2010 session track data

Query Feature	$q_{\text{m-1}}$ (RL1)	q_{m} (RL2)	SH-QM: 0.3 $q_{\text{m-1}}$ + 0.7 q_{m} (RL3)
T	0.2065	0.2134	0.2490 (+14.30%)
T, O, U	0.2092	0.2211	0.2577 (+16.55%)

2.4 Relevance Feedback (RL3 and RL4)

Although SH-QM has a natural advantage of using real (although slightly outdated) user queries for estimation, the estimation may be rough because of the limited size of query texts. Thus, in RL3 and RL4 stage, we combine pseudo-relevance feedback query model (PRF-QM) with SH-QM with the expectation of improving SH-QM. We only combine single term feature from PRF-QM and SH-QM, and keep feature O and U unchanged, as shown in (4).

In RL3, we use top 10 ranked Wikipedia documents from our RL2 results as pseudo-relevant documents and estimate query models using relevance model 1 (RM1) method [10], which is calculated as (5): W is the Wikipedia PRF document set and d can be each document in W; $p_{MLE}(t|d)$ is the probability of t from unsmoothed document model for d; $p_{SH}(q|d)$ is the probability of the weighted SDM features in SH-QM from smoothed document model for d. In RL3, we use only top ranked Wikipedia documents for PRF with the expectation of improving reliability of PRF documents, which is proved to be important for the performance of PRF [12]. Also, there were studies in 2010 web track successfully improved results using Wikipedia articles for PRF [13].

In RL4, we use the snippets (provided officially in session track topics) of the documents clicked by users as PRF documents (each clicked snippet is given equal weight in estimation), which can be calculated as (6): C is the set of clicked snippets and d refers to each of the snippets in C; $p_{MLE}(t|d)$ is the probability of t from the unsmoothed snippet model.

$$\begin{aligned} \hat{p}(t|q) &= (1-\mu) \cdot p_{SH}(t|q) + \mu \cdot p_{PRF}(t|q) \\ p_{RL3}(f|q) &= p_{RL4}(f|q) = p_{SH}(f|q), when\ F\ is\ O\ or\ U \end{aligned} \quad (4)$$

$$p_{PRF-RL3}(t|q) \propto \sum_{d \in W} p_{MLE}(t|d) p_{SH}(q|d) \quad (5)$$

$$p_{PRF-RL4}(t|q) \propto \sum_{d \in C} p_{MLE}(t|d) \quad (6)$$

In (4), the mixture weight μ is not tuned and set to 0.3 intuitively. Also, we intuitively use top 20 query terms from T feature in final ranking, which is also not tuned. As a result, results reported for our RL3 and RL4 runs may not indicate optimized performance.

Although both methods for RL3 and RL4 try to combine PRF-QM with SH-QM, the different sets of documents used for relevance feedback may lead to certain preference of methods. Compared with RL3, the documents used for RL4 (those clicked by users) are intuitively more reliable because click-through data may more reliably indicate users' positive feedback. However, because only clicked documents for past queries were available for feedback, they may be misleading when current query is very different from previous queries at sub-topic level. On the other hand, although documents for RL3 may not involve users' feedback, considering we tuned RL2 methods to optimize performance on current query, the PRF documents used for RL3 may better model user's current information needs.

Because we are not aware of the performance of the systems used for generating user interaction data in 2011 topics, we did not use

the snippets provided in official RL3 topic file, but directly used documents from our RL2 results for pseudo-relevance feedback. Thus the setting for our RL3 run is the same as that for RL2 run. This may also lead to difficulties for us to make fully comparison between RL3 and RL4 runs, because recent studies indicates PRF query models estimated from whole documents can be improved by considering positional information in PRF documents [14].

3. EVALUATION

3.1 SH-QM

We have partly evaluated results of SH-QM on 2010 session track data in 2.3. We find that SH-QM largely improved performance compared with methods using only current query. In 2011 session track, we find similar results (Table 2). In both "all-subtopics" and "current-only" evaluation, SH-QM improved nDCG@10 by over 10% (improvements are significant at 0.05 levels in paired t-test).

Table 2. nDCG@10 for RL1 and RL2 in 2011 session track and the number of topics that nDCG@10 has changed

Evaluation Setting	**RL1 (SDM+QL)**	**RL2 (SDM+SH-QM)**		
"all-subtopics"	0.3789	0.4281 (+12.98%, p = 0.002)		
Trends of change from RL1 to RL2		↑	−	↓
Number of topics		27/76	36/76	13/76
Average nDCG@10.RL1		0.3386	0.3560	0.5259
p(nDCG@10.RL1 for "↓" > nDCG@10.RL1 for "−"): 0.073 p(nDCG@10.RL1 for "↓" > nDCG@10.RL1 for "↑"): 0.043 p(nDCG@10.RL1 for "−" > nDCG@10.RL1 for "↑"): 0.413 tested by Welch's t-test				
Evaluation Setting	**RL1 (SDM+QL)**	**RL2 (SDM+SH-QM)**		
"current-only"	0.2679	0.2954 (+10.27%, p = 0.022)		
Trends of change from RL1 to RL2		↑	−	↓
Number of topics		21/76	39/76	16/76
Average nDCG@10.RL1		0.2804	0.2136	0.3838
p(nDCG@10.RL1 for "↑" > nDCG@10.RL1 for "−"): 0.162 p(nDCG@10.RL1 for "↓" > nDCG@10.RL1 for "↑"): 0.072 p(nDCG@10.RL1 for "↓" > nDCG@10.RL1 for "−"): 0.023 tested by Welch's t-test				
Pearson correlation between nDCG@10.RL2 – nDCG@10.RL1 for two different evaluation methods on 76 different topics: 0.774.				

In order to further investigate on the difference of performance for SH-QM on two evaluation settings, we compare results from two evaluation settings by per topic difference of nDCG@10 between RL1 and RL2. Figure 1 shows per topic difference of nDCG@10 between RL1 and RL2 in two evaluation settings. We find similar trends: in both settings, no difference of nDCG@10 can be found for about half of the topics, and the number of topics improved by SH-QM is more than the number of topics hurt. In Figure 2, per topic difference of nDCG@10 between RL1 to RL2 is charted for both evaluation settings and compared. Still, similar trends can be found on most of the topics.

We further compare average nDCG@10 of topics at RL1 stage for topics that are improved, unchanged, and hurt in RL2, which may help us understand in which cases RL2 can help or hurt retrieval. In both evaluation settings, the group of topics hurt by RL2 has significant higher average nDCG@10 than other two groups.

In both evaluation settings, SH-QM seems more likely to hurt performance if current query is already very effective (with higher nDCG@10), and to improve (or do not hurt) performance when current search query is comparatively less effective (with lower nDCG@10). However, it does not indicate SH-QM will improve the most difficult queries, or hurt the most effective ones.

We identified 17 topics in "all-subtopics" evaluation and 26 topics in "current-only" evaluation with lower than 0.05 nDCG@10 as difficult queries. For the 17 difficult in "all-subtopics" evaluation, 13 topics with nDCG@10 equal to 0 have not been improved in RL2 stage, and only 2 topics are effectively improved, for which the improvements in nDCG@10 are greater than 0.1; for the 26 difficult in "current-only" evaluation, 21 topics with nDCG@10 equal to 0 have not been improved in RL2, and only 3 topics are effectively improved. Thus, it seems SH-QM is not likely to be able to improve the most difficult queries. Topics being improved are mostly those with nDCG@10 from 0.2 to 0.5.

Two typical queries that are identified difficult but improved in RL2 stage are topic No. 12 and No. 73. For both topics, user's previous queries are effective, but current query has some errors: there is a typo in current query for topic No. 12; for topic 73, user issued a over-specified query, while effective query exists among previous session histories. For both topics, instead of saying RL2 improved search, it may be more appropriate to say RL2 saved users' extremely ineffective queries. However, in such cases, RL2 does not necessarily perform better than previous queries, and it is thus arguable whether it really improved users' search experience.

We also identified certain "easy" topics and found SH-QM will not hurt the topics. We identified 20 topics and 13 topics with nDCG@10 larger than 0.5 in two evaluation settings. More than half of the topics (in both settings) are not hurt by RL2. No topic is greatly hurt (with nDCG@10 decrease by more than 0.2). Thus, it seems clear that SH-QM will also not hurt those most effective queries.

Finally, we select several typical topics in 2011 for discussion.

The three topics improved most by SH-QM are: No. 13, No. 59, and No. 67. Topics No. 13 and No. 67 are both cases similar to the case of "saving user from ineffective queries": in No. 13, user used an under-generalized word "job" in current query, but "employ" used in past queries are effective; in topic No. 67, user tried to over-specify results by connecting "joseph steffen" with those from Wikipedia, but such page does not exist and previous queries are effective by just using "joseph steffen". Topic 59 may indicate a typical case of ineffective search behaviors that can be improved by SH-QM: both current query and previous queries are over-generalized, while the overlapped documents are relevant.

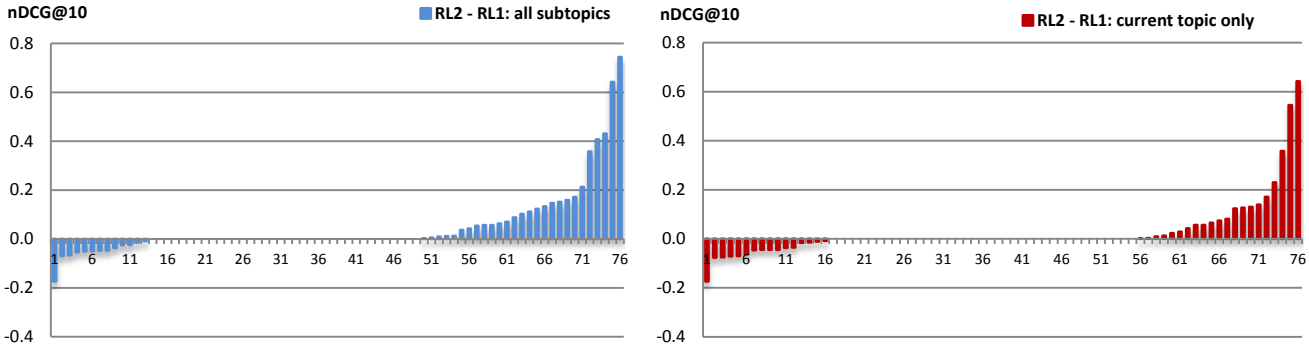

Figure 1. Difference of nDCG@10 between RL1 and RL2 on different topics (sorted by nDCG@10.RL2 – nDCG@10.RL1). Left figure indicates evaluation on all subtopics; right figure indicates evaluation on only subtopic of current query.

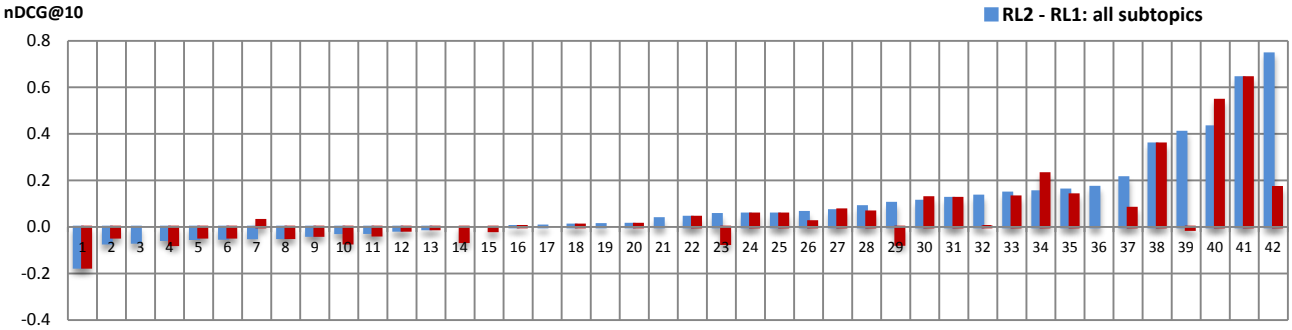

Figure 2. Comparison of nDCG@10.RL2 – nDCG@10.RL1 for two evaluation methods on different topics; only 42 topics which have difference between nDCG@10.RL2 and nDCG@10.RL1 are charted.

3.2 RL3 and RL4

Compared with the significant improvements achieved by RL2 over RL1, we do not get much improvement by combining PRF with SH-QM. For both RL3 and RL4, evaluation results on topic No. 1 – No. 46 are reported in table 3[1].

Still, we can observe significant improvements on these 46 topics between RL1 and RL2 in both evaluation settings. Compared with RL2, only RL4 is significantly better in "all-subtopics" evaluation setting (significant at 0.1 level using one tail paired t-test). But in "current-only" evaluation setting, we did not find any significant difference between RL2, RL3, and RL4. It also seems from Table 3 that click-through data (RL4) can only improve RL2 in "all-subtopics" evaluation. Comparing RL3 and RL4, in either setting, no significant difference can be claimed.

We further calculate per topic difference of nDCG@10 between RL3/RL4 and RL2. We find Pearson correlation for differences of nDCG@10 from RL2 to RL3 and that from RL2 to RL4 is -0.178 and -0.046 in two evaluation settings, which can indicate RL3 and RL4 (and possibly the different resources used for PRF) will have different but not necessarily opposite behaviors in two evaluation settings.

Table 3. nDCG@10 of RL1 to RL4 on 46 topics of 2011 session track data (topic No. 1 to No. 46)

Methods	"all-subtopics" Evaluation		"current-only" Evaluation	
RL1	0.3344		0.2080	
RL2 SH-QM	0.3811		0.2343	
	RL2:RL1	+ 13.97% p = 0.019	**RL2:RL1**	+ 12.64% p = 0.044
RL3 SH-QM + PRF using top wiki doc	0.3782		0.2371	
	RL3:RL2	− 0.76% p = 0.362	**RL3:RL2**	+ 1.20% p = 0.302
RL4 SH-QM + PRF using clicked doc	0.3993		0.2354	
	RL4:RL2	+ 4.78% p = 0.068	**RL4:RL2**	+ 0.47% p = 0.452
	RL4:RL3	+ 5.58% p = 0.072	**RL4:RL3**	− 0.72% p = 0.441

As mentioned in section 2.4, however, because related parameters are not tuned for RL3 and RL4 in our runs, results reported in this section may not indicate the optimized results for each method. Also, we do not over emphasize any conclusion in this section. However, some of the observations are likely to be generalized: RL4 and click-through data may only help RL2 in "all-subtopics" evaluation. We did not observe any improvements of using RL3. However, considering PRF on top ranked documents is usually difficult to tune, we are not going to claim PRF is not useful for RL2. Besides, according to [14], snippets used in RL4 may also contribute to the better performance of RL4 than RL3 in our runs.

[1] We made an error in generating our RL4 results submission by using our RL3 queries for topic No. 47 – No. 76 (30 topics). Thus, most results discussed in section 3.2 refer to those we only evaluated and compared for topic No. 1 – No. 46 (46 topics in total).

Table 4 reports evaluation results for our submitted runs on all topics.

Table 4. nDCG@10 and MAP of submitted results (all topics)

Evaluation Settings	Stages	nDCG@10	MAP
All Subtopics	**RL1**	0.3789	0.1206
	RL2	0.4281	0.1446
	RL3	0.4282	0.1453
	RL4	0.4409	0.1508
		nDCG@10	**MAP**
Current Subtopic Only	**RL1**	0.2679	0.1239
	RL2	0.2954	0.1391
	RL3	0.2981	0.1399
	RL4	0.2971	0.1428

4. CONCLUSION

In this paper, we introduce our methods performed in TREC 2011 session track. Our major contribution is combining different query language models: one kind of query model (SH-QM) is estimated from session historical queries; the other kind is estimated using pseudo-relevant documents. We noticed significant improvements of using SH-QM compared with ad hoc retrieval. We did not come to any solid conclusion for the benefits of combining SH-QM and PRF-QM because of some drawbacks in experiments. However, our results are most likely to support that click-through data can only significantly improve SH-QM in "all-subtopics" evaluation, but not in "current-only" evaluation. We need further experiment results to clearly find out the usefulness of two PRF resources and their benefits over SH-QM. Besides, our methods of combining different query models are extremely simple, which will be one of our major foci in future works.

5. REFERENCES

[1] Andrew Trotman and David Keeler. Ad hoc IR: not much room for improvement. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information* (SIGIR '11). ACM, New York, NY, USA, 2011: 1095-1096.

[2] http://ir.cis.udel.edu/sessions/

[3] http://lemurproject.org/clueweb09.php/

[4] Donald Metzler and W. Bruce Croft. A Markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '05). ACM, New York, NY, USA, 2005: 472-479.

[5] Gordon V. Cormack, Mark D. Smucker, and Charles L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval*, 14(5), 2011: 441-465.

[6] http://plg.uwaterloo.ca/~gvcormac/clueweb09spam/

[7] M-Dyaa Albakour, Udo Kruschwitz, Jinzhong Niu, Maria Fasli. University of Essex at the TREC 2010 Session Track. In *Proceedings of 19th Text REtrieval Conference* (TREC 2010), 2010.

[8] Sadegh Kharazmi, Falk Scholer, and Mingfang Wu. RMIT University at TREC 2010: Session Track. In *Proceedings of 19th Text REtrieval Conference* (TREC 2010), 2010.

[9] ChengXiang Zhai, John Lafferty. A risk minimization framework for information retrieval. *Information Processing & Management*, 42(1), 2006: 31-55.

[10] Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '01). ACM, New York, NY, USA, 2001: 120-127.

[11] ChengXiang Zhai and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management* (CIKM '01). ACM, New York, NY, USA, 2001: 403-410.

[12] Fernando Diaz and Donald Metzler. Improving the estimation of relevance models using large external corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '06). ACM, New York, NY, USA, 2006: 154-161.

[13] Dong Nguyen and Jamie Callan. Combination of evidence for effective web search. In *Proceedings of 19th Text REtrieval Conference* (TREC 2010), 2010.

[14] Yuanhua Lv and ChengXiang Zhai. Positional relevance model for pseudo-relevance feedback. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '10). ACM, New York, NY, USA, 2010: 579-586.