# An Investigation of Basic Retrieval Models for the Dynamic Domain Task

Razieh Rahimi and Grace Hui Yang

Department of Computer Science, Georgetown University
razieh.rahimi@georgetown.edu, huiyang@cs.georgetown.edu

**Abstract.** TREC dynamic domain is a new challenging task, which aims to simultaneously optimize the performance of retrieval and the number of iterations to accomplish the search task in a session-based search environment with a more sophisticated feedback information from the user. As a first step towards developing an effective search systems for this task, we investigate the characteristics of the newly created dataset for this task, and performance of basic well-known retrieval models for it. Our investigation demonstrates that the query sets contain multiple difficult queries, where initial results may provide very limited evidence for improvement in subsequent iterations. The new setting of the task and characteristics of dataset stress the need for more comprehensive metrics of performance evaluation, in terms of result diversity as an example.

## 1 Introduction

TREC dynamic domain (DD) task is an interactive search process where the retrieval system updates the search results based on more comprehensive feedback from a simulated user. Feedback information in this task indicates which passages of each retrieved document is related to which subtopics of the query. The goal is that the search system provides users, in the least number of iterations, with enough information regarding all aspects of a query, utilizing online feedback from users.

TREC DD task is a challenging retrieval problem due to several reasons. First of all, it is a session-based search task, which requires a more complex search system than the one with the independence assumption about queries. Second, the search system should decide to terminate the search session for a given query. Predicting whether or not the user information need is satisfied only based on partial user's feedback is difficult. Third, the granularity of search items and feedback information is different; the system searches over documents, while user's feedback specifies the relevant passages to the query. Last, receiving feedback on at most 5 documents for each query at each iteration, may lead to a limited amount of feedback information, especially for difficult queries.

We participated in TREC DD task 2016, and our main goal was to deduce properties of the task and its new dataset, that make it different from ad-hoc retrieval. Evaluation of basic retrieval models for the first iteration demonstrates that the query set of this task contains multiple difficult queries, which makes

improvement in subsequent iterations based on previously received feedback information more challenging. For subsequent iterations of retrieval, given the passage-based feedback in the DD task, we adjust the well-known relevance feedback model RM3 [6] which results in higher performance of retrieval compared to the original feedback model.

In addition, we investigated the dataset for the DD task from different points of view. Most importantly, investigation of the query set reveals that a passage relevant to a subtopic of a query usually occurs in many documents. This fact, given the difference in the granularity of retrieval and feedback segments, can cause an issue in the DD task. Recall that in the dynamic domain task, the search system provides ranked lists of documents to the user, while receives passage-based feedback from the user. And finally, the system performance is evaluated using metrics measuring the quality of (ranked) lists of documents. This setting along with the provided test dataset poses an issue, since it does not affect the performance metrics, even novelty measures, that the search system provides new information regarding an already explored subtopic to the user or provides redundant information. Therefore, search systems may tend to return redundant information instead of exploring towards new information, which is not desirable for users.

The rest of this report is organized as follows. In Section 2, retrieval models adopted for the first and subsequent iterations are described. Characteristics of the dataset for the DD task and evaluation results are discussed in Section 3. Finally, the report is concluded in Section 4.

## 2 Retrieval and Feedback Models

In this section, we describe the retrieval models used to provide results in our submitted runs to the TREC DD track. Before proceeding to the retrieval models, we first give a formal descriptions of the DD task. In the DD task, the search system receives an initial query $q$, and its goal is to satisfy the underlying information need using a given collection $C$ of documents in an interactive search process. At each iteration $i$, the system presents at most 5 documents $\{d_j^i\}_{j=1}^5$ to the user, and receives feedback in a form that indicates which passages of the presented documents are relevant to which subtopics of the query, as well as their relevance degrees. Specifically, the feedback is a set of $(p, s, r, d_j^i)$ tuples where $p$ is the passage of document $d_j^i$ that is relevant to subtopic $s$ of the query with relevance degree $r$. Based on the received feedback from iteration 1 to $i$, the search system decides to stop the search session for query $q$, or to start the new iteration $i + 1$ and present a new list of documents to the user adopting an adjusted retrieval model.

Following we describe how the results of each iteration are produced. We describe the retrieval models in two parts. First, the retrieval models used to generate the results of the first iteration is described where there is no feedback information from the (simulated) user. Incorporation of feedback information from the (simulated) user into the retrieval model is then described.

**First iteration.** In the first iteration, the document list to be shown to the user is prepared using the following methods.

*Language modeling framework.* In this approach, the top 5 documents retrieved using the language modeling framework are selected as the result of the first iteration. To rank documents using the language modeling framework, document language models are smoothed using Dirichlet prior smoothing [10].

*Relevance feedback.* In this approach, feedback information from an initial retrieval is used to expand queries. Note that the query expansion in this iteration is based on the pseudo-relevance feedback information, not precise feedback information from the (simulated) user. The top 5 documents retrieved with respect to the expanded queries are then shown to the user as the result of the first iteration.

For query expansion based on feedback documents, we use the relevance model [6]. In the first estimation method of relevance model, expansion terms are selected from the top $k$ initially retrieved documents as follows:

$$p(w|\theta_{\text{RM1}}) = \sum_{i=1}^{k} \frac{p(q|d_i)}{\mathcal{Z}} p(w|d_i), \tag{1}$$

where $p(q|d_i)$ is the retrieval score of document $d_i$ in the initial ranking for query $q$, $\mathcal{Z} = \sum_{i=1}^{k} p(q|d_i)$ is to normalize the retrieval scores, and $p(w|d_i)$ is estimated using the maximum likelihood estimate (MLE) method as $\frac{c(w,d_i)}{|d_i|}$. Initial retrieval scores $p(q|d)$ are estimated using the KL-divergence between maximum likelihood query model and document model computed using Dirichlet prior smoothing. The query language model is estimated using the MLE method as follows:

$$p(w|\theta_q) = \frac{c(w,q)}{|q|}, \tag{2}$$

where $c(w,q)$ is the count of word $w$ in query $q$ and $|q|$ shows the total number of words in the query.

In RM3 model, relevance expansion terms are linearly combined with original query terms as follows:

$$p(w|\theta_{\text{RM3}}) = \lambda p(w|\theta_q) + (1-\lambda) p(w|\theta_{\text{RM1}}), \tag{3}$$

where $\lambda \in [0,1]$ is the interpolation parameter. The documents retrieved using the $\theta_{\text{RM3}}$ query language model are used to produce the result of the first iteration.

*BM25.* In this setting, the top 5 documents ranked by the BM25 retrieval model are presented to the user.

*LDA Clustering on initial results.* In this approach, We adopt a topic modeling algorithm to diversify search results in terms of their coverage of different subtopics, as done in several studies [9]. In this regard, we first rank documents by the language modeling framework where document language models are smoothed using Dirichlet prior smoothing. We then cluster the top $k$ retrieved

documents into 5 clusters using a variant of the Latent Dirichlet Allocation (LDA) algorithm proposed in [1]. The setting of $k$ is discussed in Section 3. We opt to partition the top documents into 5 clusters, since queries of the DD track have 5 subtopics on average, as mentioned in the track description. Performing LDA on the document set, the document-cluster probability distributions are estimated. Let $p(c_i|d)$ denote the probability of cluster $i$ given document $d$, where one has $\sum_{i=1}^{5} p(c_i|d) = 1$. Based on these probability distributions, each document is assigned to the cluster that has the highest probability given the document, i.e., document $d$ is assigned to the cluster $\text{argmax}_c\, p(c|d)$. The documents in each cluster are then sorted based on their retrieval scores. Finally, the top first document from each cluster is selected to be shown to the user.

**Subsequent iterations.** In the following iteration $i > 1$, the search system has feedback information from the first $i-1$ iterations. Let us denote the set of all feedback information available prior to iteration $i$ as $F^i$ which consists of feedback tuples. We aim to obtain a diversified expansion of the original query using the feedback set $F^i$. In this regard, we try to adjust a relevance model for feedback in the language modeling retrieval framework according to the different type of feedback information in the DD task compared to ad-hoc retrieval. To adjust the relevance model for the DD task, the language model of feedback information from the Jig system is first estimated as follows:

$$p(w|\theta_{\text{JRM1}}) = \sum_{f \in F} \frac{r_f}{\mathcal{Z}} p(w|p_f), \tag{4}$$

where $p(w|p_f)$ denotes the language model of passage texts which are estimated using the MLE method, $r_f$ is the rating of the passage, and $\mathcal{Z}$ is a normalization factor for ratings. In the next step, the language model of feedback information is combined with the language model of original query as follows:

$$p(w|\theta_{\text{JRM3}}) = \lambda p(w|\theta_q) + (1-\lambda) p(w|\theta_{\text{JRM1}}), \tag{5}$$

We noticed that the passage texts in feedback information are usually short, containing one sentence where each word occurs only one time. Therefore, there is no difference between specific words and more general words in the language model of the passage text. This issue becomes acute when there are few passages in the feedback set for a query, assume the extreme case when there is only 1 passage. Thus, discrimination between specific and general words in feedback information for a query according to the language model $\theta_{\text{JRM1}}$ is usually not possible.

The issue of non-discrimination between words can be due to the fact that the inverse document frequency (IDF) effect is not considered in weighting feedback terms in relevance model for feedback information [3]. This fact has the potential to be more problematic in the setting of the DD task rather than ad-hoc retrieval. Thus, to resolve the non-discrimination issue for the DD task setting, we add an IDF-element in estimation of feedback language models as follows:

$$p(w|\theta_{\text{JRM1-IDF}}) = \sum_{f \in F^i} \frac{r_f}{\mathcal{Z}} p(w|p_f) \log \frac{|C|+2}{\text{df}(w)+1}, \tag{6}$$

**Table 1.** Results of the first iteration using different approaches.

| | CT | ACT | α-nDCG | nERR-IA | nSDCG |
|---|---|---|---|---|---|
| LM-Dirichlet | **0.2174** | **0.1516** | **0.2952** | **0.2691** | **0.1901** |
| RM3 | 0.1688 | 0.1270 | 0.2537 | 0.2408 | 0.1710 |
| LM-Dirichlet + LDA | 0.1966 | 0.1447 | 0.2692 | 0.2534 | 0.1652 |
| BM25 | 0.1818 | 0.1291 | 0.2434 | 0.2261 | 0.1564 |

where $\mathrm{df}(w)$ denotes the document frequency of term $w$, and $|C|$ shows the number of documents in the collection. Finally, this language model is interpolated with the query language model similar to Eq. 5, referred to as JRM3-IDF model. The top 5 documents retrieved with respect to the expanded query using JRM3-IDF mdoel are selected as the result of iteration $i$ to be presented to the user.

## 3 Evaluation Results

In this section, we report the results of the retrieval and feedback models described in the previous section.

**Experimental setup.** We index the collection of each domain separately using the Galago toolkit.[1] Words are stemmed using Krovetz stemmer. Queries of each domain are searched over the index of the corresponding collection, since the domains of queries are specified according to the track guideline.

**Evaluation metrics.** The results of the DD tasks, according to the track guideline, is evaluated using three categories of metrics. The first category of metrics measures how well a search system performs considering the number of iterations done to obtain the document set. This category includes Cube Test (CT) and Average Cube Test (ACT) measures proposed in [7].

The second category includes metrics to evaluate the diversification effectiveness of a document ranking, such as $\alpha$-nDCG@$k$ [2] and nERR-IA [8]. And the last category contains snDCG metric that evaluates the effectiveness of document rankings over an entire search session [5].

**Results of the first Iteration.** In Table 1, we report the performance evaluation of the results obtained for the first iteration using different retrieval models. The parameter settings for obtaining the results in this table are as follows. The parameter $\mu$ in Dirichlet prior smoothing is set to the default value of 1500. Then, for relevance feedback, the top 10 retrieved documents are used to estimate the language model $\theta_{\mathrm{RM1}}$ in Eq. 1, and the top 5 terms in this language model are used to expand the original query terms, where the combination parameter $\lambda$ in Eq. 3 is set to the default value 0.75. For the run using topic modeling, the top 200 retrieved documents are clustered. We also did not tune the parameters of the topic modeling algorithm, such as the number of iterations, and default parameter values are used. The parameters of BM25 retrieval model are also set to the default values. According to the results in Table 1, the best performance is achieved using the language modeling framework.

[1] http://www.lemurproject.org/galago.php

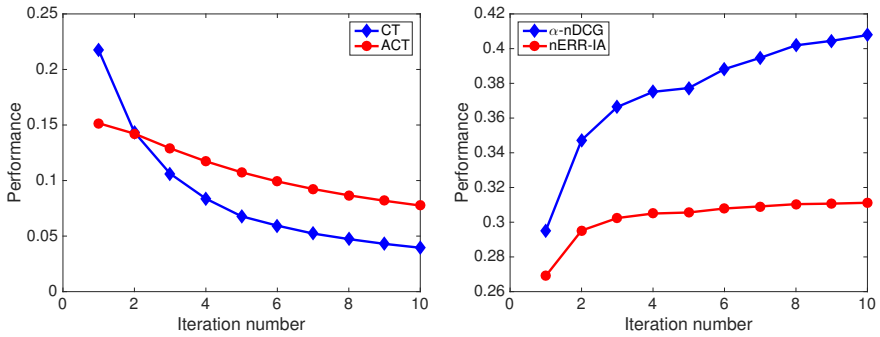
**Table 2.** Results of the second iteration using different approaches.

| | | CT | ACT | $\alpha$-nDCG | nERR-IA | nSDCG |
|---|---|---|---|---|---|---|
| 1st Iter | LM-Dirichlet | 0.2174 | 0.1516 | 0.2952 | 0.2691 | 0.1901 |
| 2nd Iter | JRM | 0.1384 | 0.1411 | 0.3340 | 0.289 | 0.1067 |
| | JRM + IDF | **0.1434** | **0.1422** | **0.3473** | **0.2952** | **0.1118** |

Since the language modeling framework achieves the highest performance, we use feedback models on this retrieval framework for later iterations, and next we report the performance of adopting feedback models for the second iteration.

**Second iteration.** Similar to the parameter settings of the original relevance model that we use for the first iteration, parameters of the JRM3 and JRM3-IDF models are not tuned, and default parameter values are used as mentioned above. The results reported in Table 2 are obtained when duplicate documents are removed from the ranked list of the second iteration, and the list contains 5 documents.[2]

The results in Table 2 show the performance of the proposed adjusted feedback models for the DD task. As reported in the table, adding IDF effect in JRM3-IDF model improves all performance measures. However, the values of CT and ACT scores are decreased in the second iteration, which is not desirable.




(a) CT and ACT scores (b) nDCG and nERR-IA scores

**Fig. 1.** Performance evaluation of results of ten iterations.

**Results of the first ten iterations.** We now report the performance of the retrieval system for the DD task where the result of first iteration is obtained by the language modeling framework, and the results of subsequent iterations are obtained by JRM3-IDF feedback model. Figure 1 shows the performance of this system for the first ten iterations. Similarly, duplicate documents in later iterations are removed, and the size of the result list at each iteration is 5. Based

[2] The results reported by the track organizers do not have duplicate documents, but for some cases the ranked lists of the second iteration does not have 5 documents.

**Table 3.** Results of ad-hoc retrieval on Ebola domain.

| MAP | P@10 |
|---|---|
| 0.2131 | 0.4519 |

on the diagram in Figure 1(a), the CT and ACT scores are decreasing as the number of iteration increases. Thus, first iteration is the best point to stop the search task according to these measures.

**Discussion.** The results of ad-hoc retrieval on ebola domain, only first iteration of interaction, are evaluated using mean average precision (MAP) and precision at the top 10 documents (P@10) metrics. The measures are calculated on the top 1,000 documents retrieved by the language modeling framework with Dirichlet prior smoothing. These evaluation results, reported in Table 3, show that the provided query set contains many difficult queries. About 50% of queries have values of average precision lower than 0.016. These results demonstrate that the search system would not get valuable feedback for many queries after first iteration, and the search system requires more sophisticated method to handle such queries.

The evaluation results for the polar domain is not provided, since some files in the collection contain duplicate copy of documents. However, this query set also seems to be a difficult query set, since the query set is a mixture of short keyword and verbose queries. Each type of queries needs to be treated differently to achieve an acceptable performance [4].

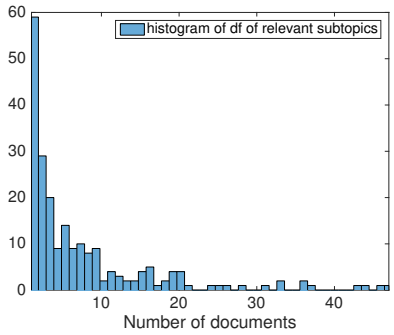



**Fig. 2.** Histogram of document frequency of a passage relevant to a subtopic.

Further investigation of dataset reveals that a passage relevant to a subtopic occurs in some documents of the collection. The diagram in Figure 2 shows the frequency of the number of different documents having the same passage relevant to a subtopic of a query, where the maximum number of documents over different passages relevant to a subtopic of a query is counted for each subtopic, and the numbers greater than 50 (about 27 subtopics, and the maximum document frequency is 3231) are removed from the diagram for clearance. The maximum and

average document frequencies of different passages relevant to 183 subtopics of 242 total subtopics are greater than 1. This characteristic of the query set along with fine-grained judgment information demonstrates that different documents presented to the user (in the same or different iterations) may contain redundant information, which is not desirable from the user perspective. The user desires new information even regarding an aspect of his/her query for which has already obtained some information. This desire can be interpreted as recall of passages relevant to each subtopic of a topic. The mentioned characteristic of the query set stresses the need for a more sophisticated metric to evaluate diversity in search results. Otherwise, a retrieval model that uses the feedback passages as a new query would probably have higher performance, than the one which uses the feedback passages for query expansion. Using the feedback passages as a new query for retrieval, in addition to provide high accuracy regarding the explored subtopics, may also provide information about new subtopics. This can happen because our investigation shows that there are 2,483 samples among 15,448 unique document-query judgment records (more than 16% of all) that their documents are relevant to more than one subtopic of the query.

Finally, comparison of different stopping strategies based on current performance metrics seems not trivial, since cube test metrics are generally decreasing with the increasing of iteration number, and diversity measures are always increasing due to evaluation on accumulated results. Therefore, the result of comparison between stopping strategies is readily obvious, the later a strategy decides to stop the search session, the lower the values of cube test metrics, the higher the values of diversity metrics.

## 4 Conclusion

In this report, we described the various retrieval models used for the dynamic domain task, and presented their results. Our evaluation shows that the query sets of the dynamic domain track contain several difficult queries, and it is not even trivial to get acceptable retrieval performance for the first iteration of interaction with the user. Therefore, improvement of search results for such queries based on the user's feedback is challenging.

In the dynamic domain task, the search system should deal with data segments of two different granularity levels; the search system provides ranked lists of documents to the user, while receives passage-based feedback from the user. And finally, the system performance is evaluated using metrics measuring the quality of (ranked) lists of documents. This setting along with the provided test dataset poses the following challenges. First, pseudo-relevance feedback techniques need to be adjusted for the setting of dynamic domain task to produce more effective expanded query, as demonstrated in the results of our experiments by adding the inverse-document frequency heuristic. Second, evaluation metrics need also to be adjusted to evaluate novelty in two levels, first covering different subtopics of a query, and then covering as diverse information as possible regard-

ing each subtopic. This 2-level result diversification is possible in the setting of the dynamic domain task.

**Acknowledgments.** This research was supported by DARPA grant FA8750-14-2-0226, NSF grant IIS-145374, and NSF grant CNS-1223825. Any opinions, findings, conclusions, or recommendations expressed in this paper are of the authors, and do not necessarily reflect those of the sponsor.

## References

1. S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 280–288. JMLR: W\&CP 28 (2), 2013.
2. C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 659–666, New York, NY, USA, 2008. ACM.
3. S. Clinchant and E. Gaussier. A theoretical analysis of pseudo-relevance feedback models. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, ICTIR '13, pages 6:6–6:13, New York, NY, USA, 2013. ACM.
4. M. Gupta and M. Bendersky. Information retrieval with verbose queries. *Foundations and Trends in Information Retrieval*, 9(3-4):209–354, 2015.
5. K. Järvelin, S. L. Price, L. M. L. Delcambre, and M. L. Nielsen. Discounted cumulated gain based evaluation of multiple-query ir sessions. In *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*, ECIR'08, pages 4–15, Berlin, Heidelberg, 2008. Springer-Verlag.
6. V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 120–127, New York, NY, USA, 2001. ACM.
7. J. Luo, C. Wing, H. Yang, and M. Hearst. The water filling model and the cube test: Multi-dimensional evaluation for professional search. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 709–714, New York, NY, USA, 2013. ACM.
8. T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1043–1052, New York, NY, USA, 2011. ACM.
9. R. L. T. Santos, C. Macdonald, and I. Ounis. Search result diversification. *Found. Trends Inf. Retr.*, 9(1):1–90, Mar. 2015.
10. C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 334–342, New York, NY, USA, 2001. ACM.