TREC 2016 Dynamic Domain Track Overview

Grace Hui Yang[1] and Ian Soboroff[2]

[1] Department of Computer Science
Georgetown University
huiyang@cs.georgetown.edu
[2] NIST
ian.soboroff@nist.gov

1 Introduction

The main goal of TREC dynamic domain track is to explore and evaluate systems for interactive information retrieval, which reflects real-world search scenarios. Due to the importance of learning from user interactions, this track has been held for the second year. The name of the track contains two parts. "Dynamic" means that the search system dynamically adapts the provided ranking to the user through interactions. "Domain" stems from the fact that the search task in the track is on the domains of special interests, which tend to bring information needs that would not be met within a single interaction. The task is inspired by interested groups in government, including the DARPA Memex program.[3]

Each search task in the DD track involves some interactions between a user and a search system. In the first iteration, the user submits a query and the target domain of interest to the search system. The system provides the user with an initial ranking of documents, and receives feedback from the user on the provided ranking. This interaction, providing the user with a ranking and receiving the user's feedback, continues until the search system stops the search task. The DD track introduces a new challenging search problem with the following important assumptions.

1. The DD task is an interactive search task, where the search system needs to dynamically adapt its ranking based on users' feedback.
2. During the search session, the user does not provide a new formulation of his/her information need.
3. The user provides fine-grained feedback information on a received list of documents, in which the passages of the retrieved documents that are relevant to his/her information need are specified. In addition, user's feedback indicates that passages of interest are relevant to which subtopic of the query with graded relevance degrees.
4. The search system is required to stop the search task when the user is provided with enough information regarding all aspects of his/her information need.

[3] The DARPA Memex program aims to advance the state of the art in domain-specific web crawling, visualization, and discovery.

These new settings of the search task in the DD track are motivated by professional search tasks such as finding a criminal network, or finding prior art for a patent application. In addition, the DD track emphasizes on finer-grained relevance judgments compared to open domain Web searches, since professional users have stringent relevancy requirements best expressed at the passage level rather than the whole document level.

To simulate the above defined search process, the DD track provided two newly-created corpora on specific Ebola and Polar domains. In addition, a system that simulates users' feedback based on fine-grained relevance judgment information provided for the query sets, is developed specifically for the task.

A total of 6 groups participated in the track this year, a slight decrease from last year, when 7 groups participated. One submitted run by the group from the RMIT university is marked as a manual run by the participants, which provides an interesting baseline for the task.

2 Dynamic Domain Track

2.1 Dynamic Domain Task Description

The dynamic domain task simulates an interactive search process, where the search system tries to dynamically adapt its ranking based on the user's feedback obtained in previous interactions.

At the first iteration, the search system receives an initial query for each topic. In response to that query, the system may present up to five documents to the user. The system then receives feedback only on the presented documents in a form that indicates which passages of the retrieved documents are relevant to the topic, and specifically to which subtopic of the topic along with the graded relevance degree. The DD Track emphasizes on finer-grained relevance judgment because professional users have stringent relevancy requirements best expressed at the passage level rather than the whole document level. Note that although feedback information provides relevance status regarding subtopics of the query, the total number of subtopics of a topic is not given to the search system in the beginning of a search session nor during the interactions with the user through feedback. Receiving feedback from the user, the search system decides whether to stop the interaction, or continues by providing the user an updated list of at most 5 documents. The two possible actions of the search system shows two points. The first point is that the search system is supposed to stop search sessions when sufficient information regarding all subtopics of a query is presented to the user. Based on this, one evaluation metric measures the speed of providing relevant documents to the user. The second point is that the search system needs to adopt received feedback in previous iterations to provide more effective results to the user in the next iteration.

The users' feedback is simulated by a system called Jig.[4] The search systems interact with Jig, and receive feedback on retrieved documents at each iteration.

[4] The implementation of the Jig system is available online at https://github.com/trec-dd/trec-dd-jig.

Table 1. Dataset statistics.

Domain	Size of data on disk	Number of documents	Number of queries
Ebola	12.6 GB	682,157	27
Polar	158 GB	1,741,530	26

The Jig system provides relevance information in the form described above only when relevance information on given documents is available. In contrast to the traditional TREC interpretation of judgments, no feedback on a document does not mean that the document is irrelevant to the query. This setting makes using of negative feedback challenging.

The participants were supposed to develop an adapting search system. The Track looks forward to systems that are able to make educated guesses for later queries based on early feedback from the user, so that the entire search process can be sped up. Further, the Track expects the search system, not the user, to decide when to stop the search. This requires the search systems to just provide the right amount of information.

Listing 1.1. One sample topic

```
<topic name=''US Military Crisis Response'' id=''DD16-1'' num_of_subtopics=''3''
    >
    <subtopic name=''West African mission'' id=''DD16-1.1'' num_of_passages
        =''1862''>
    </subtopic>
    <subtopic name=''Key Personnel'' id=''DD16-1.2'' num_of_passages=''1091''>
    </subtopic>
    <subtopic name=''Personnel safety protocols'' id=''DD16-1.3'' num_of_passages
        =''692''>
    </subtopic>
</topic>
```

Listing 1.2. Sample judgment

```
<subtopic name=''West African mission'' id=''DD16-1.1'' num_of_passages=''1862''
    >
            <passage id=''25268''>
                <docno>ebola-634445
                    eda14aa2756fbd3eff24b0ccf10f24543c4da9bdd54cbb354c46ba5c66</
                    docno>
                <rating>3</rating>
                <text><![CDATA[In this week's AFRICOM Update engineers continue
                    to build Ebola Treatment Units in Liberia while a special
                    facility for infected healthcare workers nears completion.]]
                    ></text>
                <type>MANUAL</type>
            </passage>
            <passage id=''25273''>
                <docno>ebola-87627477
                    a4223a52b2caa93f3a80637b9760191187ac7a27942f16c62e2fb52c</
                    docno>
                <rating>3</rating>
                <text><![CDATA[this week's AFRICOM Update engineers continue to
                    build Ebola Treatment Units in Liberia while a special
                    facility for infected healthcare workers nears completion.]]
                    ></text>
                <type>MATCHED</type>
                <score>0.93</score>
            </passage>
            ...
</subtopic>
```

2.2 Datasets and Domains

The TREC DD track provides two domains of documents; Ebola and Polar. Table 1 reports the statistics of the two domains. All the datasets are format-

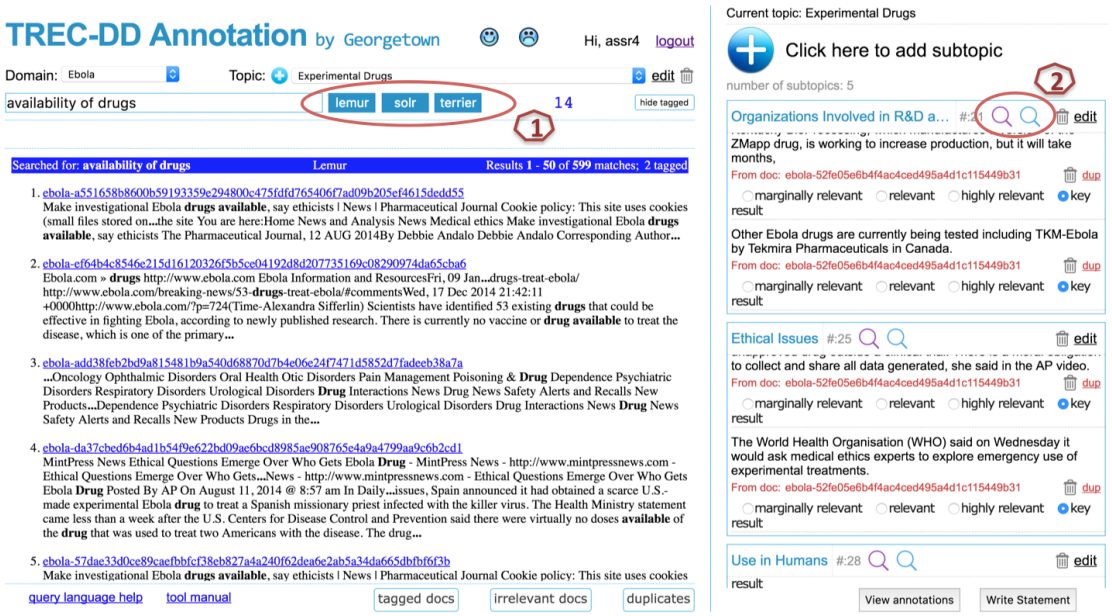

Fig. 1. TREC 2016 DD Track Annotation Tool.

ted using the Common Crawl Architecture schema from the DARPA MEMEX project, and stored as sequences of CBOR objects. To find more details about the datasets, please refer to [1]. There are 53 topics in total. One sample of topics is shown in Listing 1.1. For each topic, there is relevance judgment information on the passage level as shown in Listing 1.2. Following, the process of assessing documents and passages is described.

Topic and Assessment Development The topics were developed by six NIST assessors over five weeks in the spring of 2016. A topic (which is like a query) contains a few words. It is the main search target for the dynamic search process. Each topic contains multiple subtopics, each of which addresses one aspect of the topic. Each subtopic contains multiple relevant passages that the assessors discovered from across the entire corpus. Each passage is tagged with a grade to mark how relevant it is to the subtopic. We treat the obtained set of passages as the complete set of relevant passages and use them in the evaluation.

The NIST assessors were asked to produce a complete set of subtopics for each topic using an annotation tool (Fig. 1). The tool provides five retrieval algorithms to compensate each other, which include the default search algorithms in Lemur[5]

[5] `http://www.lemurproject.org/`

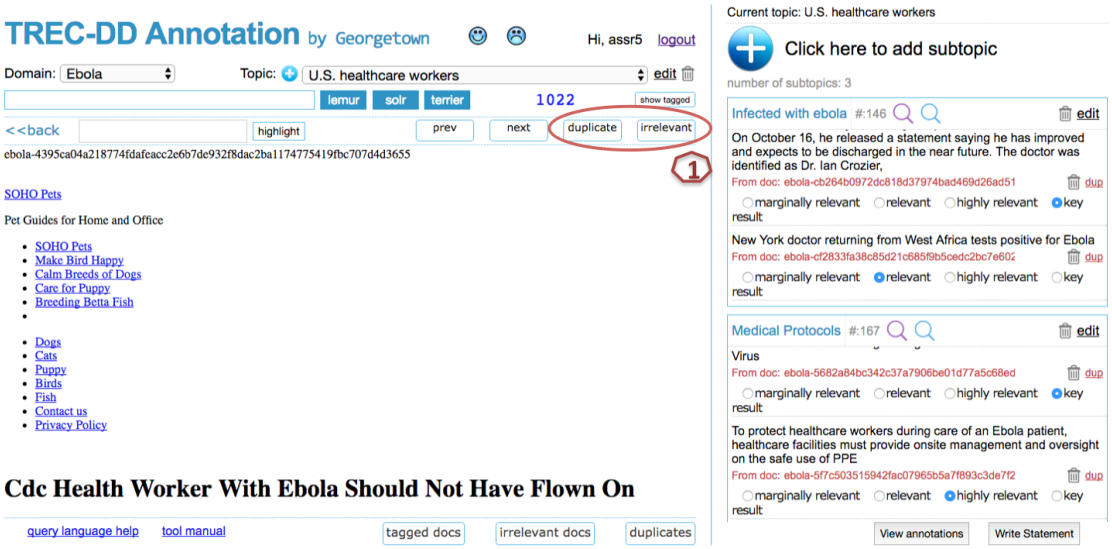

Fig. 2. TREC 2016 DD Track Annotation Tool.

, Solr[6] and Terrier[7] (Circle 1 in Fig. 1), one relevance feedback algorithm and one adaptive search algorithm (Circle 2 in Fig. 1).

To get a list of documents to examine, the assessors enter search queries and choose Lemur, Solr or Terrier to retrieve. While examining the documents returned by the search engine, they can drag and drop a text fragment of any length to a box to mark it as relevant to a subtopic. Then they can grade the text fragments at a scale of 1: marginally relevant, 2:relevant, 3:highly relevant, and 4:key results. The assessors can also mark a document as irrelevant or duplicate to the topics (Circle 1 in Fig 2).

The assessors can also get document list via relevance feedback algorithm and adaptive search algorithm, which are the pink and blue magnifiers in Fig. 1. The relevance feedback algorithm utilizes a subtopic's title and all relevant context to the subtopic to expand the original search query. The adaptive search algorithm utilize both relevant and irrelevant contents to generate retrieval list. Notice that our relevance feedback and adaptive search algorithms are conducted at the subtopic level, which means only relevant/irrelevant texts to the corresponding subtopics are utilized to optimize retrieval results.

For example (Fig. 1), "assr4" first used Lemur to search for query "availability of drugs". Along the process of examining retrieval results, s/he generated subtopics "Organizations involved in R&D . . . ", "Ethical Issues", "Use in Humans" etc. for the topic "Experimental Drugs" and also located some relevant texts for these passages. And then s/he could generate more documents to exam

[6] `https://lucene.apache.org/solr/`
[7] `http://terrier.org/`

by using Solr or Terrier to retrieve for the same query. S/he could also try to find more documents related to subtopic "Ethical Issues" by clicking the pink/blue magnifier icon next to the corresponding subtopic.

2.3 Evaluation Measures

The primary measures to evaluate the effectiveness of search systems in the DD track are Average Cube Test (ACT) and Cube Test (CT), proposed in [5]. Both measures evaluate the speed of completion of an entire search task; how fast a system could fill up the task cube with diverse and relevant information. Before the detailed description of evaluation measures, we introduce some notations. We denote a query by q, the set of its subtopics by $\mathcal{S}_q$, a single document by d, and a set of documents by D.

The first measure CT is defined as follows:

$$\mathrm{CT}(q, D) = \frac{\mathrm{Gain}(q, D)}{\mathrm{Time}(D)}, \tag{1}$$

where $\mathrm{Time}(D)$ denotes the number of iterations to obtain document set D, and $\mathrm{Gain}(q, D)$ is estimated by the following formula.

$$\mathrm{Gain}(q, d_j) = \sum_{s \in \mathcal{S}_q} \Gamma \theta_s \mathrm{rel}(d_j, s) \mathbb{1} \left(\sum_{k=1}^{j-1} \mathrm{rel}(d_k, s) < \mathrm{MaxHeight} \right), \tag{2}$$

where the elements of the formula are as follows:

- Γ is a discounting factor to include novelty in the evaluation and is calculated by
$$\Gamma = \gamma^{\mathrm{nrel}(s, D^{j-1})}, \tag{3}$$
where γ denotes the discount factor, and $\mathrm{nrel}(s, D^{j-1})$ is the number of documents relevant to subtopic s in the set of documents ranked higher than document d_j.
- θ_s denotes the importance degree of subtopic s such that one has $\sum_{s \in \mathcal{S}_q} \theta_i = 1$,
- $\mathbb{1}$ is the indicator function,
- rel() denotes the relevance degree between a document and a subtopic, calculated as an average over all its passages relevant to the subtopic, and it is in the $[0, 1]$ range.

The second metric for evaluation of search systems considering the time taken to accomplish the search task is Average Cube Test (ACT) defined as an average of values of cube test metric, calculated at each rank in the list. In particular, ACT metric is calculated as follows:

$$\mathrm{ACT}(q, D) = \frac{1}{|D|} \sum_{k=1}^{|D|} \frac{\mathrm{Gain}(\mathrm{q}, \mathrm{D^k})}{\mathrm{Time}(D^k)}, \tag{4}$$

Table 2. Participant groups.

Group	Country
Federal University of Minas Gerais (UFMG)	Brazil
Georgetown University	USA
Laval University & Lakehead University	Canada
NanJing University (IAPLab)	China
RMIT University	Australia
University of Padua (UPD_IA)	Italy

where D^k is the set of documents from rank 1 to rank k.

Finally, in the TREC 2016 DD, the parameters of these two evaluation metric are set as: $\gamma = 0.5$, MaxHeight $= 5$, and all subtopics are assumed as equally important.

In addition, since the search tasks are all multiple-faceted, we include IR metrics measuring subtopic relevance, including α-nDCG@k [3] and nERR-IA [2]. We also consider evaluation using session-based measures, such as snDCG [4]. At each iteration, the value of these measures are computed on the ranked list of documents obtained by concatenating all ranked lists from the first iteration to that iteration.

3 Submission and Results

We received 21 submissions from 6 groups mentioned in Table 2.

Summary of the adopted methods. Descriptions of the submitted runs provided by the participant group are as follows:

Run-ID:rmit-oracle.lm.1000: We run Solr with the content language model to get the first 1000 documents, then we use the ground truth to remove non relevant documents from the initial list of documents. For each iteration, we return the next 5 relevant documents from the initial list. A document is relevant if it was found in the topic's list of judged documents. The motive is to estimate an upper bound of the task and understand if the first 1000 documents are enough to get all relevant documents.

Run-ID:rmit-lm-rocchio.Rp.NRd.10: We use the content of documents to build a content language model and get the top 5 documents. We then use the Rocchio algorithm to reformulate the current iteration query using the feedback provided by JIG. To represent relevant documents, we concatenate relevant passages from relevant documents into a pseudo relevant passages (Rp) whereas we use the content of the non relevant documents as the non relevant units of Rocchio (NRd). Lastly, we use the top 10 non negative terms from the new query vector generated by Rocchio to build the new query. In addition, we set Rocchio parameters to α=1, β=0.75 and γ=0.25.

rmit_lm_nqe: In this method, we used the Language modeling approach as implemented in Apache Solr using Dirichlet smoothing and default parameters. We leveraged Solr's edismax query parser that scores documents by the similarity

score between the page content and the sum of bi-gram and uni-gram queries. No query expansion (nqe) was applied.

ufmgXS2: Flat diversification with single-source subtopics and cumulative stopping condition.

ufmgHS2: Hierarchical diversification with single-source subtopics and cumulative stopping condition.

ufmgHM3: Hierarchical diversification with multi-source subtopics and window-based stopping condition.

ufmgHM2: Hierarchical diversification with multi-source subtopics and cumulative stopping condition.

ufmgXM2: Flat diversification with multi-source subtopics and cumulative stopping condition.

LDA_Indri73: We use Indri and LDA to access the first iteration of the first query. Then we use the MDP model which we has modified and get the next iteration.During the MDP model,we use the Indri to help to search and then get the final result.

rmit_lm_psg.max: We split documents into half overlapped passages with a passage size of 200 words and index them as documents alongside their parent documents in Apache Solr. We then use Solr's block join query to score documents based on the maximum of their passage level relevance scores. The method scores passages using the sum of the passage language model score for a unigram query and a bigram based phrase query.

UL_LDA_NE: LDA used on a corpus of 25 documents from solr, Oriented for NE topics by reducing the text from each document to sentences which contains a part of the NE.

UPD_IA_BiQBFi: BM25 followed by 5 iterations of feedback based on an algorithm inspired by Quantum Detection (QB) that exploits binary representation for documents. Feedback consists in re-ranking the (residual) top 1000 documents. When relevant documents are present in the feedback set explicit feedback is performed; when no relevant documents are present, residual collection is re-ranked by PRF on the top 100 documents. Description selection is based on WPQ; top 35 terms + topic terms are used.

UPD_IA_BiQBDiJ: BM25 followed by max 5 iterations of feedback based on an algorithm inspired by Quantum Detection (QB) that exploits binary representation for documents. Feedback consists in re-ranking the (residual) top 1000 documents. When relevant documents are present in the feedback set explicit feedback is performed; when no relevant documents are present, residual collection is re-ranked by PRF on the top 100 documents. Description selection is based on WPQ; top 35 terms + topic terms are used. After two PRF-based re-ranking no additional iterations are performed.

UL_BM25: BM25 similarity.

UL_Kmeans: Kmeans applied on a subset of documents retrieved by Solr, best document of each cluster is returned to the user.

UL_LDA_200: LDA is used to create 5 different topics from documents. We take 100 results from Solr, we remove documents which are too similar to others

documents, then we fill the dataset with other documents to have 100 document to run LDA over the sample.

UL_LDA_Psum: Probability for each document to be assigned to each topic multiplied by the global probability of each topic to obtain the document which covers the maximum of topic information.

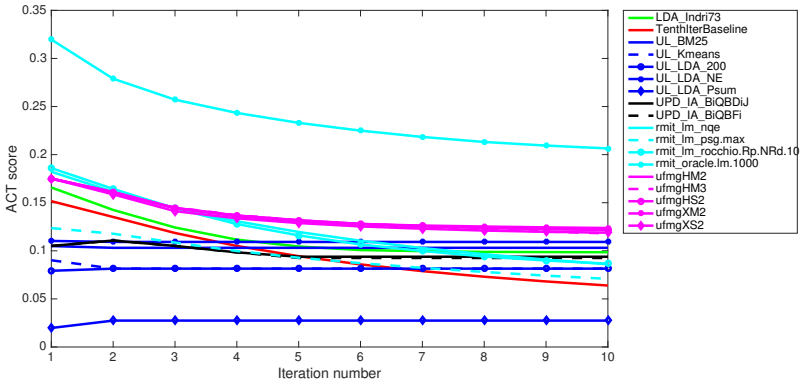

Fig. 3. ACT scores of submitted runs over ten iterations.

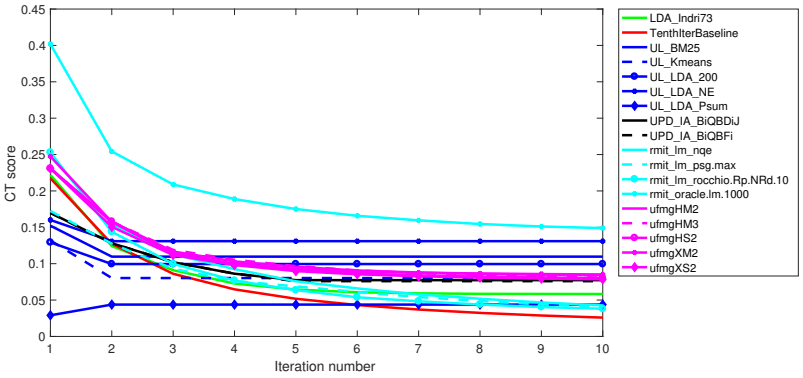

Fig. 4. CT scores of submitted runs over ten iterations.

Results. Figures 3 and 4 show the change of ACT and CT scores, respectively, over ten iterations for all runs. The evaluation results of the cube test and average cube test measures show a decreasing trend with iteration for almost all submitted runs. We further provide the change of ACT score for two specific

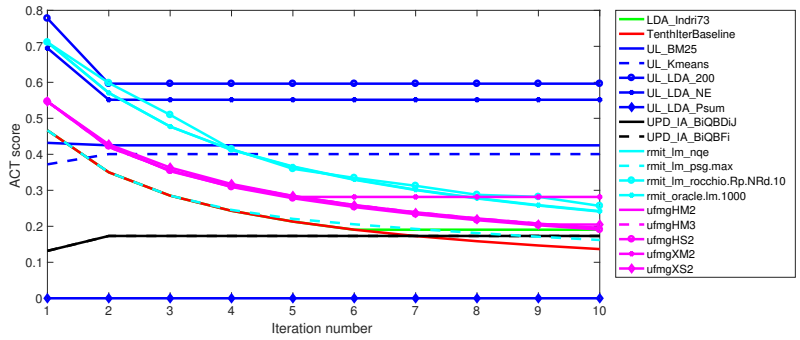

Fig. 5. ACT scores of query DD16-20 over ten iterations.

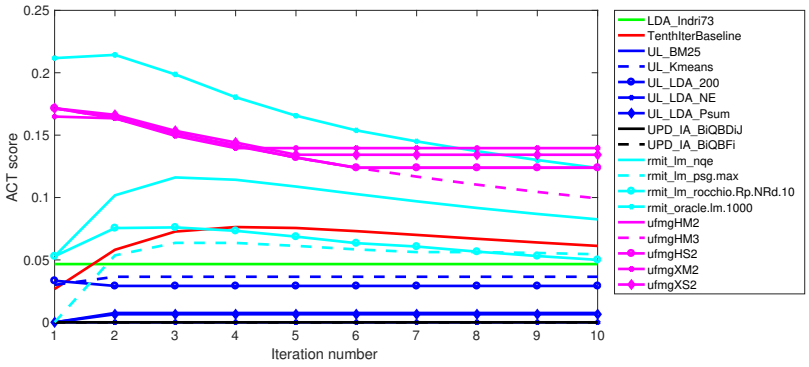

Fig. 6. ACT scores of query DD16-16 over ten iterations.

queries, DD16-20 and DD16-16. These DD16-20 and DD16-16 queries have 1.0 and 0.0 values for precision at 5 documents measure, respectively, thus they seem to be samples of difficult and easy queries. Figures 5 and 6 show the change of ACT scores for queries DD16-20 and DD16-16, respectively. The ACT scores for query DD16-16 are lower than those for query DD16-20, however some runs have improved the ACT score for query DD16-16 in following iterations, which is not the case for query DD16-20.

4 Conclusion

The dynamic domain track ran for the second time at TREC 2016, focusing on the significant interactive search task based on the newly-built dataset and designed settings.

We received 21 runs from 6 groups. The evaluation results demonstrate that the distance between the manual run and the best automatic run is substantial, therefore the dynamic domain task is a difficult search task and further investigation is required to achieve acceptable performance. In addition, the decision to stop the search session requires estimation of user satisfaction which is very challenging, and is rarely addressed in received runs.

Acknowledgments. The TREC 2016 Dynamic Domain Track is sponsored by the DARPA Memex program. We have our special thanks to Razieh Rahimi, Jiyun Luo, Yunyun Chen, Shiqi Liu from Georgetown university, and Shahzad Rajput from NIST. We thank the following contributors to TREC DD Track in crawling the data: Diffeo, Giant Oak, Hyperion Gray, NASA JPL, and New York University.

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. datasets for dd track. `http://trec-dd.org/dataset.html`.
2. O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S.-L. Wu. Intent-based diversification of web search results: Metrics and algorithms. *Inf. Retr.*, 14(6):572–592, Dec. 2011.
3. 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.
4. 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.

5. 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.

Table 3. TREC 2016 Dynamic Domain Track Evaluation Results - Iteration 1

Run ID	ACT	CT	α-nDCG	nERR-IA	AVG-αnDCG	AVG-nERRIA	nSDCG	Precision
rmit_oracle.lm.1000	0.3196	0.4019	0.6874	0.6709	0.0568	0.0549	0.4993	0.9623
rmit_lm_rocchio.Rp.NRd.10	0.1860	0.2530	0.3715	0.3547	0.0312	0.0271	0.2531	0.3890
rmit_lm_nqe	0.1822	0.2479	0.3581	0.3438	0.0309	0.0268	0.2470	0.3708
ufmgXS2	0.1751	0.2309	0.3516	0.3383	0.0289	0.0280	0.2226	0.4000
ufmgHS2	0.1751	0.2309	0.3516	0.3383	0.0289	0.0280	0.2226	0.4000
ufmgHM3	0.1751	0.2309	0.3516	0.3383	0.0289	0.0280	0.2226	0.4000
ufmgHM2	0.1751	0.2309	0.3516	0.3383	0.0289	0.0280	0.2226	0.4000
ufmgXM2	0.1750	0.2474	0.3559	0.3355	0.0384	0.0371	0.2254	0.4226
LDA_Indri73	0.1658	0.2220	0.3288	0.3150	0.0312	0.0272	0.2166	0.3811
TenthIterBaseline	0.1516	0.2174	0.2952	0.2691	0.0274	0.0231	0.1901	0.3208
SecondIterationBaseline	0.1516	0.2174	0.2952	0.2691	0.0274	0.0231	0.1901	0.3208
FirstIterBaseline	0.1516	0.2174	0.2952	0.2691	0.0274	0.0231	0.1901	0.3208
FifthIterBaseline	0.1516	0.2174	0.2952	0.2691	0.0274	0.0231	0.1901	0.3208
rmit_lm_psg.max	0.1236	0.1725	0.2707	0.2611	0.0193	0.0177	0.1655	0.3179
UL_LDA_NE	0.1103	0.1601	0.2257	0.2073	0.0306	0.0305	0.0954	0.2906
UPD_IA_BiQBFi	0.1050	0.1698	0.2275	0.2011	0.0176	0.0151	0.1412	0.2453
UPD_IA_BiQBDiJ	0.1050	0.1698	0.2275	0.2011	0.0176	0.0151	0.1412	0.2453
UL_BM25	0.1050	0.1522	0.2094	0.1919	0.0161	0.0148	0.1178	0.2340
UL_Kmeans	0.0902	0.1319	0.1796	0.1630	0.0136	0.0124	0.0799	0.2340
UL_LDA_200	0.0792	0.1293	0.1740	0.1593	0.0183	0.0163	0.0709	0.2075
UL_LDA_Psum	0.0197	0.0291	0.0558	0.0514	0.0043	0.0038	0.0154	0.0868

Detailed Results

The evaluation scores for the submitted runs calculated for iteration 1 to 10 are listed in Tables 3 to 12, respectively. The average of the scores over all the topics are reported when duplicate documents in subsequent ranked lists are removed.

Table 4. TREC 2016 Dynamic Domain Track Evaluation Results - Iteration 2

Run ID	ACT	CT	α-nDCG	nERR-IA	AVG-αnDCG	AVG-nERRIA	nSDCG	Precision
rmit_oracle.lm.1000	0.2789	0.2543	0.7116	0.6830	0.0310	0.0293	0.3102	0.9623
rmit_lm_rocchio.Rp.NRd.10	0.1644	0.1439	0.3883	0.3630	0.0174	0.0144	0.1236	0.3422
rmit_lm_nqe	0.1614	0.1539	0.3908	0.3611	0.0164	0.0139	0.1243	0.3393
ufmgXM2	0.1612	0.1574	0.4028	0.3588	0.0217	0.0198	0.1275	0.4038
ufmgHS2	0.1611	0.1581	0.4079	0.3664	0.0182	0.0158	0.1305	0.4075
ufmgHM3	0.1601	0.1578	0.4055	0.3653	0.0183	0.0159	0.1306	0.4075
ufmgHM2	0.1601	0.1578	0.4055	0.3653	0.0183	0.0159	0.1306	0.4075
ufmgXS2	0.1587	0.1506	0.3987	0.3622	0.0165	0.0150	0.1305	0.4283
FirstIterBaseline	0.1516	0.2174	0.2952	0.2691	0.0274	0.0231	0.1901	0.3208
LDA_Indri73	0.1425	0.1242	0.3350	0.3172	0.0162	0.0139	0.1000	0.2547
TenthIterBaseline	0.1352	0.1274	0.3142	0.2777	0.0160	0.0127	0.0920	0.2528
SecondIterationBaseline	0.1352	0.1274	0.3142	0.2777	0.0160	0.0127	0.0920	0.2528
FifthIterBaseline	0.1352	0.1274	0.3142	0.2777	0.0160	0.0127	0.0920	0.2528
rmit_lm_psg.max	0.1178	0.1275	0.3166	0.2836	0.0116	0.0098	0.0940	0.2985
UPD_IA_BiQBFi	0.1107	0.1281	0.2736	0.2238	0.0105	0.0086	0.0909	0.3038
UPD_IA_BiQBDiJ	0.1107	0.1281	0.2736	0.2238	0.0105	0.0086	0.0909	0.3038
UL_LDA_NE	0.1092	0.1309	0.2779	0.2319	0.0192	0.0177	0.0703	0.2868
UL_BM25	0.1031	0.1097	0.2520	0.2131	0.0098	0.0083	0.0759	0.2811
UL_Kmeans	0.0815	0.0803	0.1922	0.1685	0.0072	0.0064	0.0386	0.1887
UL_LDA_200	0.0815	0.0995	0.2110	0.1772	0.0121	0.0096	0.0431	0.1981
UL_LDA_Psum	0.0274	0.0438	0.1039	0.0750	0.0052	0.0034	0.0165	0.1811

Table 5. TREC 2016 Dynamic Domain Track Evaluation Results - Iteration 3

Run ID	ACT	CT	α-nDCG	nERR-IA	AVG-αnDCG	AVG-nERRIA	nSDCG	Precision
rmit_oracle.lm.1000	0.2571	0.2089	0.7251	0.6879	0.0215	0.0201	0.2616	0.9623
FirstIterBaseline	0.1516	0.2174	0.2952	0.2691	0.0274	0.0231	0.1901	0.3208
ufmgHM3	0.1450	0.1190	0.4306	0.3742	0.0131	0.0109	0.0932	0.3950
ufmgHM2	0.1447	0.1167	0.4254	0.3725	0.0135	0.0113	0.0928	0.3899
ufmgHS2	0.1445	0.1142	0.4256	0.3728	0.0131	0.0109	0.0895	0.3899
rmit_lm_nqe	0.1444	0.1129	0.4073	0.3674	0.0114	0.0094	0.0886	0.3244
rmit_lm_rocchio.Rp.NRd.10	0.1442	0.0998	0.3931	0.3647	0.0121	0.0100	0.0814	0.3184
ufmgXM2	0.1441	0.1134	0.4176	0.3642	0.0151	0.0134	0.0868	0.3774
ufmgXS2	0.1412	0.1119	0.4161	0.3683	0.0122	0.0104	0.0910	0.4113
SecondIterationBaseline	0.1352	0.1274	0.3142	0.2777	0.0160	0.0127	0.0920	0.2528
LDA_Indri73	0.1242	0.0911	0.3435	0.3201	0.0115	0.0098	0.0675	0.2119
TenthIterBaseline	0.1185	0.0861	0.3145	0.2780	0.0107	0.0084	0.0579	0.1786
FifthIterBaseline	0.1185	0.0861	0.3145	0.2780	0.0107	0.0084	0.0579	0.1786
UL_LDA_NE	0.1092	0.1309	0.2779	0.2319	0.0192	0.0177	0.0703	0.2868
rmit_lm_psg.max	0.1086	0.0930	0.3321	0.2890	0.0080	0.0066	0.0643	0.2787
UPD_IA_BiQBFi	0.1051	0.1020	0.2922	0.2307	0.0075	0.0059	0.0718	0.3220
UPD_IA_BiQBDiJ	0.1051	0.1020	0.2922	0.2307	0.0075	0.0059	0.0718	0.3220
UL_BM25	0.1031	0.1097	0.2520	0.2131	0.0098	0.0083	0.0759	0.2811
UL_Kmeans	0.0815	0.0803	0.1922	0.1685	0.0072	0.0064	0.0386	0.1887
UL_LDA_200	0.0815	0.0995	0.2110	0.1772	0.0121	0.0096	0.0431	0.1981
UL_LDA_Psum	0.0274	0.0438	0.1039	0.0750	0.0052	0.0034	0.0165	0.1811

Table 6. TREC 2016 Dynamic Domain Track Evaluation Results - Iteration 4

Run ID	ACT	CT	α-nDCG	nERR-IA	AVG-αnDCG	AVG-nERRIA	nSDCG	Precision
rmit_oracle.lm.1000	0.2434	0.1889	0.7346	0.6907	0.0171	0.0158	0.2445	0.9623
FirstIterBaseline	0.1516	0.2174	0.2952	0.2691	0.0274	0.0231	0.1901	0.3208
ufmgHM3	0.1371	0.1054	0.4377	0.3763	0.0106	0.0088	0.0858	0.3884
ufmgHM2	0.1367	0.1032	0.4325	0.3745	0.0108	0.0090	0.0853	0.3833
ufmgHS2	0.1366	0.0989	0.4335	0.3750	0.0107	0.0088	0.0810	0.3846
ufmgXM2	0.1360	0.1001	0.4243	0.3662	0.0120	0.0106	0.0774	0.3651
SecondIterationBaseline	0.1352	0.1274	0.3142	0.2777	0.0160	0.0127	0.0920	0.2528
ufmgXS2	0.1338	0.0981	0.4262	0.3711	0.0104	0.0086	0.0820	0.4025
rmit_lm_nqe	0.1307	0.0925	0.4250	0.3726	0.0088	0.0071	0.0688	0.3170
rmit_lm_rocchio.Rp.NRd.10	0.1278	0.0768	0.3968	0.3658	0.0094	0.0078	0.0621	0.2953
LDA_Indri73	0.1116	0.0727	0.3457	0.3206	0.0089	0.0076	0.0526	0.1840
UL_LDA_NE	0.1092	0.1309	0.2779	0.2319	0.0192	0.0177	0.0703	0.2868
TenthIterBaseline	0.1051	0.0647	0.3141	0.2779	0.0080	0.0063	0.0420	0.1358
FifthIterBaseline	0.1051	0.0647	0.3141	0.2779	0.0080	0.0063	0.0420	0.1358
UL_BM25	0.1031	0.1097	0.2520	0.2131	0.0098	0.0083	0.0759	0.2811
rmit_lm_psg.max	0.0998	0.0761	0.3493	0.2941	0.0062	0.0050	0.0491	0.2696
UPD_IA_BiQBFi	0.0984	0.0863	0.3016	0.2335	0.0058	0.0046	0.0603	0.3204
UPD_IA_BiQBDiJ	0.0984	0.0862	0.3009	0.2333	0.0064	0.0051	0.0602	0.3195
UL_Kmeans	0.0815	0.0803	0.1922	0.1685	0.0072	0.0064	0.0386	0.1887
UL_LDA_200	0.0815	0.0995	0.2110	0.1772	0.0121	0.0096	0.0431	0.1981
UL_LDA_Psum	0.0274	0.0438	0.1039	0.0750	0.0052	0.0034	0.0165	0.1811

Table 7. TREC 2016 Dynamic Domain Track Evaluation Results - Iteration 5

Run ID	ACT	CT	α-nDCG	nERR-IA	AVG-αnDCG	AVG-nERRIA	nSDCG	Precision
rmit_oracle.lm.1000	0.2330	0.1751	0.7387	0.6917	0.0142	0.0130	0.2347	0.9623
FirstIterBaseline	0.1516	0.2174	0.2952	0.2691	0.0274	0.0231	0.1901	0.3208
SecondIterationBaseline	0.1352	0.1274	0.3142	0.2777	0.0160	0.0127	0.0920	0.2528
ufmgHM2	0.1318	0.0957	0.4348	0.3750	0.0092	0.0076	0.0817	0.3784
ufmgHM3	0.1317	0.0967	0.4417	0.3773	0.0089	0.0074	0.0807	0.3810
ufmgHS2	0.1316	0.0910	0.4360	0.3756	0.0092	0.0075	0.0781	0.3819
ufmgXM2	0.1305	0.0927	0.4258	0.3666	0.0102	0.0089	0.0743	0.3562
ufmgXS2	0.1289	0.0898	0.4280	0.3715	0.0089	0.0073	0.0789	0.3967
rmit_lm_nqe	0.1196	0.0758	0.4322	0.3745	0.0072	0.0057	0.0556	0.2940
rmit_lm_rocchio.Rp.NRd.10	0.1160	0.0636	0.4010	0.3669	0.0077	0.0063	0.0506	0.2790
UL_LDA_NE	0.1092	0.1309	0.2779	0.2319	0.0192	0.0177	0.0703	0.2868
LDA_Indri73	0.1045	0.0645	0.3480	0.3212	0.0074	0.0063	0.0456	0.1683
UL_BM25	0.1031	0.1097	0.2520	0.2131	0.0098	0.0083	0.0759	0.2811
TenthIterBaseline	0.0944	0.0518	0.3138	0.2778	0.0064	0.0051	0.0327	0.1087
FifthIterBaseline	0.0944	0.0518	0.3138	0.2778	0.0064	0.0051	0.0327	0.1087
UPD_IA_BiQBDiJ	0.0938	0.0774	0.3066	0.2348	0.0058	0.0046	0.0543	0.3072
rmit_lm_psg.max	0.0928	0.0688	0.3675	0.2983	0.0051	0.0040	0.0415	0.2706
UPD_IA_BiQBFi	0.0925	0.0760	0.3073	0.2350	0.0047	0.0037	0.0529	0.3059
UL_Kmeans	0.0815	0.0803	0.1922	0.1685	0.0072	0.0064	0.0386	0.1887
UL_LDA_200	0.0815	0.0995	0.2110	0.1772	0.0121	0.0096	0.0431	0.1981
UL_LDA_Psum	0.0274	0.0438	0.1039	0.0750	0.0052	0.0034	0.0165	0.1811

Table 8. TREC 2016 Dynamic Domain Track Evaluation Results - Iteration 6

Run ID	ACT	CT	α-nDCG	nERR-IA	AVG-αnDCG	AVG-nERRIA	nSDCG	Precision
rmit_oracle.lm.1000	0.2250	0.1658	0.7406	0.6921	0.0121	0.0111	0.2286	0.9623
FirstIterBaseline	0.1516	0.2174	0.2952	0.2691	0.0274	0.0231	0.1901	0.3208
SecondIterationBaseline	0.1352	0.1274	0.3142	0.2777	0.0160	0.0127	0.0920	0.2528
ufmgHM2	0.1281	0.0907	0.4363	0.3753	0.0081	0.0067	0.0796	0.3772
ufmgHS2	0.1278	0.0864	0.4366	0.3757	0.0086	0.0070	0.0766	0.3809
ufmgXM2	0.1278	0.0892	0.4259	0.3666	0.0098	0.0086	0.0727	0.3536
ufmgHM3	0.1276	0.0908	0.4439	0.3777	0.0078	0.0064	0.0781	0.3794
ufmgXS2	0.1254	0.0854	0.4295	0.3719	0.0078	0.0064	0.0776	0.3967
rmit_lm_nqe	0.1105	0.0660	0.4419	0.3766	0.0061	0.0048	0.0474	0.2813
UL_LDA_NE	0.1092	0.1309	0.2779	0.2319	0.0192	0.0177	0.0703	0.2868
rmit_lm_rocchio.Rp.NRd.10	0.1066	0.0538	0.4018	0.3671	0.0065	0.0054	0.0422	0.2614
UL_BM25	0.1031	0.1097	0.2520	0.2131	0.0098	0.0083	0.0759	0.2811
LDA_Indri73	0.1008	0.0607	0.3526	0.3222	0.0065	0.0056	0.0435	0.1618
FifthIterBaseline	0.0944	0.0518	0.3138	0.2778	0.0064	0.0051	0.0327	0.1087
UPD_IA_BiQBDiJ	0.0938	0.0774	0.3066	0.2348	0.0058	0.0046	0.0543	0.3072
UPD_IA_BiQBFi	0.0925	0.0760	0.3073	0.2350	0.0047	0.0037	0.0529	0.3059
rmit_lm_psg.max	0.0872	0.0611	0.3764	0.3001	0.0047	0.0035	0.0350	0.2664
TenthIterBaseline	0.0859	0.0431	0.3137	0.2778	0.0053	0.0042	0.0268	0.0906
UL_Kmeans	0.0815	0.0803	0.1922	0.1685	0.0072	0.0064	0.0386	0.1887
UL_LDA_200	0.0815	0.0995	0.2110	0.1772	0.0121	0.0096	0.0431	0.1981
UL_LDA_Psum	0.0274	0.0438	0.1039	0.0750	0.0052	0.0034	0.0165	0.1811

Table 9. TREC 2016 Dynamic Domain Track Evaluation Results - Iteration 7

Run ID	ACT	CT	α-nDCG	nERR-IA	AVG-αnDCG	AVG-nERRIA	nSDCG	Precision
rmit_oracle.lm.1000	0.2184	0.1596	0.7438	0.6928	0.0107	0.0098	0.2244	0.9623
FirstIterBaseline	0.1516	0.2174	0.2952	0.2691	0.0274	0.0231	0.1901	0.3208
SecondIterationBaseline	0.1352	0.1274	0.3142	0.2777	0.0160	0.0127	0.0920	0.2528
ufmgXM2	0.1263	0.0876	0.4265	0.3667	0.0097	0.0085	0.0723	0.3532
ufmgHS2	0.1257	0.0840	0.4367	0.3757	0.0082	0.0066	0.0758	0.3804
ufmgHM2	0.1256	0.0879	0.4367	0.3754	0.0073	0.0060	0.0786	0.3764
ufmgHM3	0.1246	0.0871	0.4446	0.3779	0.0069	0.0057	0.0768	0.3779
ufmgXS2	0.1230	0.0827	0.4313	0.3722	0.0071	0.0058	0.0769	0.3951
UL_LDA_NE	0.1092	0.1309	0.2779	0.2319	0.0192	0.0177	0.0703	0.2868
UL_BM25	0.1031	0.1097	0.2520	0.2131	0.0098	0.0083	0.0759	0.2811
rmit_lm_nqe	0.1029	0.0577	0.4461	0.3774	0.0053	0.0041	0.0407	0.2679
rmit_lm_rocchio.Rp.NRd.10	0.0998	0.0483	0.4050	0.3678	0.0058	0.0048	0.0372	0.2532
LDA_Indri73	0.0997	0.0592	0.3532	0.3223	0.0065	0.0055	0.0425	0.1578
FifthIterBaseline	0.0944	0.0518	0.3138	0.2778	0.0064	0.0051	0.0327	0.1087
UPD_IA_BiQBDiJ	0.0938	0.0774	0.3066	0.2348	0.0058	0.0046	0.0543	0.3072
UPD_IA_BiQBFi	0.0925	0.0760	0.3073	0.2350	0.0047	0.0037	0.0529	0.3059
rmit_lm_psg.max	0.0823	0.0542	0.3821	0.3011	0.0041	0.0030	0.0307	0.2612
UL_Kmeans	0.0815	0.0803	0.1922	0.1685	0.0072	0.0064	0.0386	0.1887
UL_LDA_200	0.0815	0.0995	0.2110	0.1772	0.0121	0.0096	0.0431	0.1981
TenthIterBaseline	0.0789	0.0370	0.3137	0.2778	0.0046	0.0036	0.0226	0.0776
UL_LDA_Psum	0.0274	0.0438	0.1039	0.0750	0.0052	0.0034	0.0165	0.1811

Table 10. TREC 2016 Dynamic Domain Track Evaluation Results - Iteration 8

Run ID	ACT	CT	α-nDCG	nERR-IA	AVG-αnDCG	AVG-nERRIA	nSDCG	Precision
rmit_oracle.lm.1000	0.2132	0.1545	0.7448	0.6929	0.0095	0.0087	0.2209	0.9623
FirstIterBaseline	0.1516	0.2174	0.2952	0.2691	0.0274	0.0231	0.1901	0.3208
SecondIterationBaseline	0.1352	0.1274	0.3142	0.2777	0.0160	0.0127	0.0920	0.2528
ufmgXM2	0.1253	0.0866	0.4265	0.3667	0.0096	0.0084	0.0721	0.3523
ufmgHM2	0.1238	0.0859	0.4367	0.3754	0.0067	0.0056	0.0780	0.3739
ufmgHS2	0.1238	0.0822	0.4367	0.3757	0.0078	0.0063	0.0753	0.3800
ufmgHM3	0.1221	0.0845	0.4458	0.3781	0.0063	0.0052	0.0756	0.3725
ufmgXS2	0.1211	0.0808	0.4323	0.3723	0.0065	0.0053	0.0761	0.3914
UL_LDA_NE	0.1092	0.1309	0.2779	0.2319	0.0192	0.0177	0.0703	0.2868
UL_BM25	0.1031	0.1097	0.2520	0.2131	0.0098	0.0083	0.0759	0.2811
LDA_Indri73	0.0989	0.0586	0.3532	0.3223	0.0064	0.0055	0.0421	0.1565
rmit_lm_nqe	0.0964	0.0518	0.4515	0.3784	0.0047	0.0036	0.0359	0.2626
FifthIterBaseline	0.0944	0.0518	0.3138	0.2778	0.0064	0.0051	0.0327	0.1087
rmit_lm_rocchio.Rp.NRd.10	0.0940	0.0439	0.4073	0.3682	0.0052	0.0043	0.0335	0.2459
UPD_IA_BiQBDiJ	0.0938	0.0774	0.3066	0.2348	0.0058	0.0046	0.0543	0.3072
UPD_IA_BiQBFi	0.0925	0.0760	0.3073	0.2350	0.0047	0.0037	0.0529	0.3059
UL_Kmeans	0.0815	0.0803	0.1922	0.1685	0.0072	0.0064	0.0386	0.1887
UL_LDA_200	0.0815	0.0995	0.2110	0.1772	0.0121	0.0096	0.0431	0.1981
rmit_lm_psg.max	0.0780	0.0487	0.3876	0.3021	0.0037	0.0026	0.0273	0.2506
TenthIterBaseline	0.0731	0.0324	0.3136	0.2778	0.0040	0.0032	0.0196	0.0679
UL_LDA_Psum	0.0274	0.0438	0.1039	0.0750	0.0052	0.0034	0.0165	0.1811

Table 11. TREC 2016 Dynamic Domain Track Evaluation Results - Iteration 9

Run ID	ACT	CT	α-nDCG	nERR-IA	AVG-αnDCG	AVG-nERRIA	nSDCG	Precision
rmit_oracle.lm.1000	0.2094	0.1511	0.7452	0.6930	0.0086	0.0079	0.2186	0.9623
FirstIterBaseline	0.1516	0.2174	0.2952	0.2691	0.0274	0.0231	0.1901	0.3208
SecondIterationBaseline	0.1352	0.1274	0.3142	0.2777	0.0160	0.0127	0.0920	0.2528
ufmgXM2	0.1244	0.0858	0.4265	0.3667	0.0095	0.0084	0.0717	0.3511
ufmgHM2	0.1228	0.0850	0.4367	0.3754	0.0066	0.0055	0.0776	0.3730
ufmgHS2	0.1226	0.0810	0.4367	0.3757	0.0077	0.0062	0.0751	0.3799
ufmgHM3	0.1200	0.0824	0.4468	0.3782	0.0058	0.0047	0.0746	0.3688
ufmgXS2	0.1198	0.0795	0.4323	0.3723	0.0062	0.0051	0.0756	0.3884
UL_LDA_NE	0.1092	0.1309	0.2779	0.2319	0.0192	0.0177	0.0703	0.2868
UL_BM25	0.1031	0.1097	0.2520	0.2131	0.0098	0.0083	0.0759	0.2811
LDA_Indri73	0.0987	0.0584	0.3532	0.3223	0.0064	0.0054	0.0421	0.1561
FifthIterBaseline	0.0944	0.0518	0.3138	0.2778	0.0064	0.0051	0.0327	0.1087
UPD_IA_BiQBDiJ	0.0938	0.0774	0.3066	0.2348	0.0058	0.0046	0.0543	0.3072
UPD_IA_BiQBFi	0.0925	0.0760	0.3073	0.2350	0.0047	0.0037	0.0529	0.3059
rmit_lm_nqe	0.0909	0.0467	0.4555	0.3790	0.0042	0.0032	0.0320	0.2546
rmit_lm_rocchio.Rp.NRd.10	0.0899	0.0404	0.4086	0.3685	0.0047	0.0039	0.0308	0.2381
UL_Kmeans	0.0815	0.0803	0.1922	0.1685	0.0072	0.0064	0.0386	0.1887
UL_LDA_200	0.0815	0.0995	0.2110	0.1772	0.0121	0.0096	0.0431	0.1981
rmit_lm_psg.max	0.0742	0.0440	0.3896	0.3023	0.0033	0.0023	0.0245	0.2391
TenthIterBaseline	0.0681	0.0288	0.3136	0.2778	0.0035	0.0028	0.0172	0.0604
UL_LDA_Psum	0.0274	0.0438	0.1039	0.0750	0.0052	0.0034	0.0165	0.1811

Table 12. TREC 2016 Dynamic Domain Track Evaluation Results - Iteration 10

Run ID	ACT	CT	α-nDCG	nERR-IA	AVG-αnDCG	AVG-nERRIA	nSDCG	Precision
rmit_oracle.lm.1000	0.2065	0.1489	0.7468	0.6932	0.0079	0.0073	0.2173	0.9623
FirstIterBaseline	0.1516	0.2174	0.2952	0.2691	0.0274	0.0231	0.1901	0.3208
SecondIterationBaseline	0.1352	0.1274	0.3142	0.2777	0.0160	0.0127	0.0920	0.2528
ufmgXM2	0.1237	0.0852	0.4265	0.3667	0.0094	0.0083	0.0714	0.3504
ufmgHM2	0.1219	0.0842	0.4367	0.3754	0.0066	0.0055	0.0774	0.3716
ufmgHS2	0.1215	0.0801	0.4368	0.3758	0.0076	0.0062	0.0749	0.3794
ufmgXS2	0.1188	0.0786	0.4324	0.3723	0.0059	0.0048	0.0752	0.3871
ufmgHM3	0.1181	0.0808	0.4481	0.3784	0.0053	0.0044	0.0737	0.3654
UL_LDA_NE	0.1092	0.1309	0.2779	0.2319	0.0192	0.0177	0.0703	0.2868
UL_BM25	0.1031	0.1097	0.2520	0.2131	0.0098	0.0083	0.0759	0.2811
LDA_Indri73	0.0985	0.0583	0.3532	0.3223	0.0064	0.0054	0.0421	0.1557
FifthIterBaseline	0.0944	0.0518	0.3138	0.2778	0.0064	0.0051	0.0327	0.1087
UPD_IA_BiQBDiJ	0.0938	0.0774	0.3066	0.2348	0.0058	0.0046	0.0543	0.3072
UPD_IA_BiQBFi	0.0925	0.0760	0.3073	0.2350	0.0047	0.0037	0.0529	0.3059
rmit_lm_rocchio.Rp.NRd.10	0.0866	0.0381	0.4097	0.3687	0.0043	0.0035	0.0293	0.2342
rmit_lm_nqe	0.0860	0.0426	0.4584	0.3794	0.0038	0.0029	0.0288	0.2484
UL_Kmeans	0.0815	0.0803	0.1922	0.1685	0.0072	0.0064	0.0386	0.1887
UL_LDA_200	0.0815	0.0995	0.2110	0.1772	0.0121	0.0096	0.0431	0.1981
rmit_lm_psg.max	0.0708	0.0404	0.3930	0.3028	0.0030	0.0021	0.0223	0.2310
TenthIterBaseline	0.0639	0.0259	0.3136	0.2778	0.0032	0.0025	0.0154	0.0543
UL_LDA_Psum	0.0274	0.0438	0.1039	0.0750	0.0052	0.0034	0.0165	0.1811