TREC 2015 Dynamic Domain Track Overview

Hui Yang
Department of Computer Science
Georgetown University
huiyang@cs.georgetown.edu

John Frank
Diffeo
MIT
jrf@diffeo.com

Ian Soboroff
NIST
ian.soboroff@nist.gov

Abstract

Search tasks for professional searchers, such as law enforcement agencies, police officers, and patent examiners, are often more complex than open domain Web search tasks. When professional searchers look for relevant information, it is often the case that they need to go through multiple iterations of searches to interact with a system. The Dynamic Domain Track supports research in dynamic, exploratory search within complex information domains. By providing real-time fine-grained feedback with relevance judgments that was collected during assessing time to the participating systems, we create a dynamic and iterative search process that lasts until the system decides to stop. The search systems are expected to be able to adjust their retrieval algorithms based on the feedback and find quickly relevant information for a multi-faceted information need. This document reports the task, datasets, topic and assessment creation, submissions, and evaluation results for the TREC 2015 Dynamic Domain (DD) Track.

1 Introduction

Professional search tasks, such as finding a criminal network or finding prior art for a patent application, are often more complex than open domain Web searches. With current query-driven search engines, searchers need to go through multiple iterations of ad-hoc searches to accomplish a complex information seeking task. The search engine is forced to interpret the session across multiple queries, clicks, and dwells. The user could provide richer feedback if the system had a stronger notion of the complex task driving the individual interactions. State-of-the-art search systems have worked extensively on one-shot keyword searches where users find a single answer and quickly exit the search, yet little is known about how search systems react over the entire search process of completing a task. Inspired by interested groups in government, including the

DARPA Memex program[1], we proposed a new track in 2015 called the Dynamic Domain Track, to bring corpora, tasks, and evaluation to dynamic search in complex information domains.

The Dynamic Domain (DD) Track is interested in studying and evaluating the entire information seeking process when a search engine is dynamically adapting to a dedicated search user's feedback. The name of the track contains two parts. "Dynamic" means that it is a dynamic search task. "Domain" means that the Track is from domains of special interests that usually produce complex search tasks which would not be accomplished with a single search.

In TREC 2015, the DD Track simulates a dynamic and interactive search process. The participating systems interact with a simulated user (called "the jig") to get real-time feedback on a short list of documents the systems return. The search systems then adapt their retrieval algorithms to return a new list of search results and get another iteration of feedback. The process repeats until the search stops. The search task is assumed to have a finishing line to reach, and the task is expected to be finished as soon as possible, rather than just finding more relevant information in a recall-oriented task. 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. In addition, the DD Track emphasizes on finer-grained relevance judgments because professional users have stringent relevancy requirements best expressed at the passage level rather than the whole document level. Various evaluation metrics used in the Track are designed to measure the effectiveness of search systems in dynamic search.

The following sections report the task, datasets, topic and assessment creation, submissions, and evaluation results for the Track.

2 Domains and Corpora

In TREC 2015, the DD Track has provided and used the following datasets.

- Ebola. This data is related to the Ebola outbreak in Africa in 2014-2015. The dataset comprises 497,362 web pages, 19,834 PDFs, and 164,961 tweets. The web pages primarily contain information from NGO's, relief agencies, and news organizations. The PDF documents come from West African government and other sources. The tweets subset includes tweets that originate from West African regions involved in the Ebola outbreak. The aim of this dataset is to provide information to citizens and aid workers on the ground. The total size of the data is 12.6 GB.

- Illicit Goods. This dataset is related to how fraudulent accounts, fake product reviews, link farms, and other forms of "black hat SEO" are

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

Table 1: TREC 2015 DD Dataset Statistics.

Dataset	#docs	size	avg doc length	#unique terms
DD Ebola	0.66M	13GB	0.95K	1.1M
DD Illicit Goods	0.47M	8GB	0.46K	3.6M
DD Local Politics	0.96M	58GB	1.21K	1.1M

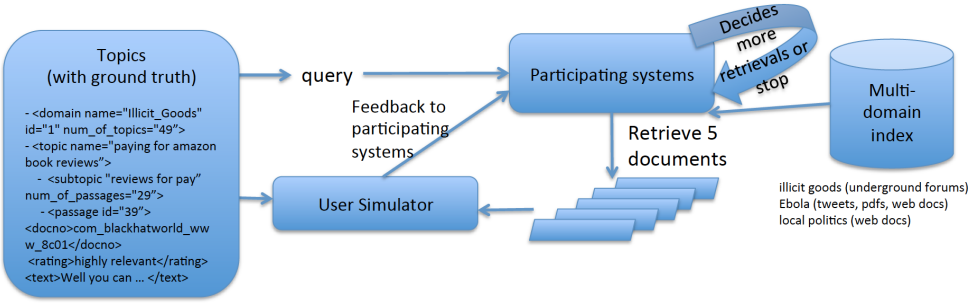

Figure 1: TREC DD Task Illustration.

made, advertised, and sold on the Internet. The dataset, whose size is 8.3GB, comprises 526,717 threads (3,345,133 posts) and contains 3.5 million unique terms. These threads are from underground hacking forums, BlackHatWorld.com and HackForums.com, with each record containing the HTML of the thread, extracted posts and metadata.

- Local Politics. This dataset is a subset of the TREC 2014 KBA Stream Corpus, whose size is 58GB, containing 6,831,397 web pages and about 1.1 million unique terms. It is related to regional politics in the Pacific Northwest and the small-town politicians and personalities that are a part of that. This dataset has HTML web news from many sources, collected as part of the KBA 2014 Stream Corpus. The HTML has been cleansed using the Boilerpipe extraction system, so the content should limited to the content of the news item.

The documents are stored in files that each contain a stream of CBOR records. CBOR is a variation of JSON that supports binary data and has more efficient encoding than text. To get access to the collections, you should follow instruction posted on `http://trec-dd.org/dataset.html`. Table 1 summaries the statistics of TREC 2015 DD Track.

3 Task

In 2015, the TREC DD Track has a simple information seeking model which is based on a notion of a professional searcher driving a system by feedback rather than queries. The participating systems receive an initial query for each topic,

where the query is two to four words and additionally indicates the domain. In response to that query, systems may return up to five documents to the user. The simulated user responds by indicating which of the retrieved documents are relevant to their interests in the topic, and to which subtopic the document is relevant to. Additionally, the simulated user identifies passages from the relevant documents and assigns the passages to the subtopics with a graded relevance rating. The system may then return another five documents for more feedback. The retrieval loop continues until the system decides to stop. All the interactions, aka, the multiple cycles of retrieval results are used to evaluate a system's search effectiveness. An effective participating system is expected to be able to find as many of the relevant documents as possible, using fewer interactions. Fig. 1 illustrate the TREC 2015 DD Track task.

The ground truth passages are used as the feedback to interact with the search engine. Below is an example of a fragment of interaction history:

- Topic DD-51: "Theresa Spence"
- System retrieves 5 documents,
- User responds that document #3 is relevant (grade "4") to some subtopic ("32"):
  feedback:[[], [], ["1323453660-374c2bc4b4371a227d4b9ff703c9750e", "32", "My community will not consider third party managers nor pay for them out of our already depressed band support funding budget, Attawapiskat First Nation Chief Theresa Spence wrote, mostly in capital letters, in a response to Aboriginal Affairs Minister John Duncan on Friday. ", "4"], [], []]
- System retrieves 5 documents,
- User responds that none are relevant:
  feedback:[[], [], [], [], []]
- System retrieves 5 documents
- User indicates a relevant passage from document #4:
  feedback:[[], [], [], [["1323124200-1f699d3ee9a338089fa0bc6ec6t1h2eb1t7o3p", "32", "The government said earlier it had chosen Jacques Marion, from the accounting and consulting firm BDO Canada, as its third-party manager for Attawapiskat. Marion was to exercise signing authority for all department spending and would decide which band staff are required to run its program and services. Spence said the minister responsible for First Nations didn't listen. We/d like to work together but put third party away. We've demonstrated we have our deficit down. We don't need a banker to come and tell us what to do, the chief told Solomon. ", "4"]], []]
- System retrieves 5 documents

- User responds that the system has found information relevant to a different subtopic ("70").
  feedback:[[["1322812020-35e06badddb1bd58ca16a34dfffedef3","70", "Roughly $11 million in debt, the Attawapiskat council has been on financial alert since at least July 2010, when federal officials required it to engage a professional co-manager to monitor spending and accounting procedures. The council hired Clayton Kennedy, who admits he is in a romantic relationship with Spence. Yes, I am, he said Thursday. There is a conflict-of-interest policy in place. Kennedy said he does not make recommendations to the chief or her deputy but to the council as a whole, and so his personal relationship with Spence should not be a factor.", "4"]], [], [], [], []]

A software tool simulating the user is provided to show instant relevance feedback to participating systems. The package, called the "jig" runs on Linux, Mac OS, and Windows.[2] The jig always provides feedback for every result, even if the feedback is that the system has no truth data for that result. At each iteration, the evaluation metric scores (Section 5) are also provided through the jig to the participating systems. By simulating the user via the jig, we enforce the interaction model and limit the exposure of relevant information to the system.

4 Topic and Assessment Development

The topics were developed by six NIST assessors over five weeks in the spring of 2015. 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 a tool powered by Lemur[3] and an active learning tool. To get a list of documents to examine, the assessors entered search queries into the search engine or fetched documents via the active learning tool. While examining the documents returned either by the search engine or by the active learning tool, the assessors could drag and drop a text fragment of any length to a box to mark it as relevant to a subtopic. This feedback was delivered to the active learning backend to generate a dynamic "frontier" of documents to assess. The assessors could also grade the text fragments at a scale of 1: marginally relevant, 2:relevant, 3:highly relevant, and 4:key results. The assessors could view the status of any document in the collection. These statuses include a

[2]For TREC 2015 DD, the jig package can be found at https://github.com/trec-dd/trec-dd-simulation-harness.

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

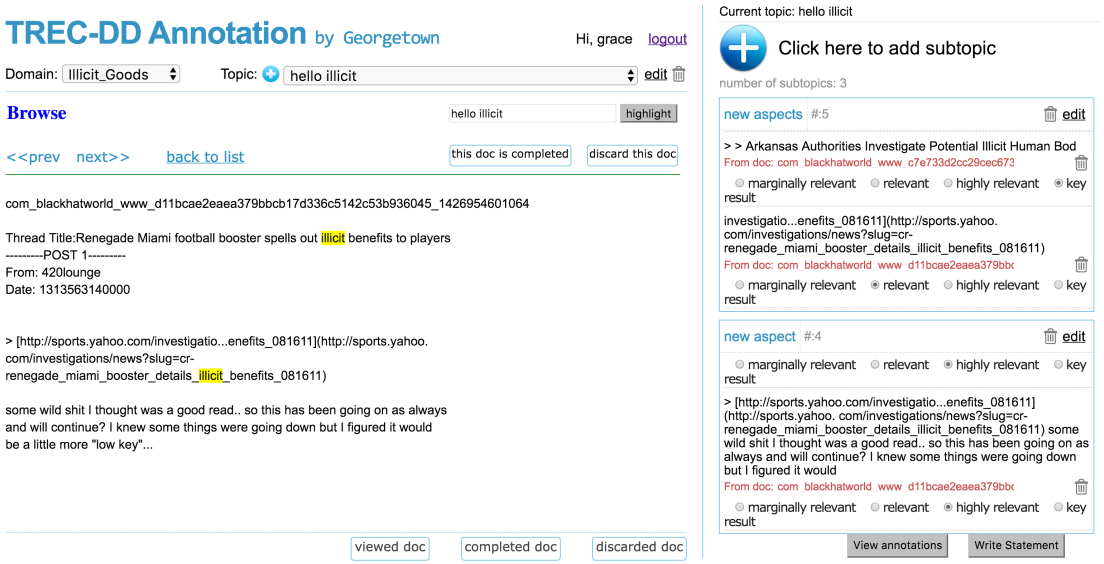

Figure 2: TREC 2015 DD Track Annotation Tool.

Table 2: TREC 2015 DD Topic Statistics.

Dataset	#topics	#subtopic per topic	avg. # subtopics covered by a rel. doc	avg rel. docs per subtopic	avg rel. docs per topic
DD Ebola	40	5.7	1.3	136	603
DD Illicit Goods	30	5.3	1.3	9	39
DD Local Politics	48	5.5	1.6	42	141

document being deleted, viewed, and annotated relevant where at least one text fragment was dragged and dropped from it.

In total, the assessors created 118 topics in the DD Track: 48 for Local Politics, 40 for domain Ebola, and 30 for domain Illicit Goods. In total, 58,758 relevant passages were found. The maximum number of of relevant passages for a topic is 8,672 and the minimum is 3. The average number of relevant passages per topic is 498. Table 2 summarizes the topic statistics.

5 Evaluation Metrics

An ideal complex search evaluation metric would measure how well a search system allows the user to handle the trade-offs between time taken to search and how well the returned documents cover the different aspects of the information need. We pick metrics which could be able to evaluate the effectiveness of the entire process of dynamic search. The main measure used in the DD Track is the Cube Test [3]. It has two variations, Cube Test (CT) and Average Cube

Test (ACT). Both are used in TREC 2015 DD evaluation.

- Cube Test (CT) is computed at the end of each search iteration. It is calculated as
$$CT(Q, D) = Gain(Q, D)/Time(D) \tag{1}$$
- Averaged Cube Test (ACT) is calculated at each document being retrieved. It is calculated as:
$$AVG\_CT(Q, D) = \frac{1}{|D|} \sum_{k} \frac{Gain(Q, D^k)}{Time(D^k)} \tag{2}$$
k is the k^{th} document in D. Basically we calculate the CT score for each document d in D, then calculate the mean value of these CT scores.

where $Gain(Q, D)$ is the accumulated gain of relevance documents in the document set D. For a list of documents D, $Gain(Q, D) = \sum_{d_j \in D} Gain(Q, d_j)$. $Gain(Q, d_j)$ is calculated as

$$\text{Gain}(Q, d_j) = \sum_{i} \Gamma \theta_i rel(d_j, c_i) \mathcal{I} \left(\sum_{k=1}^{j-1} rel(d_k, c_i) < \text{MaxHeight} \right) \tag{3}$$

where $rel()$, a score in [0,1], denotes the relevance grade between a document and a subtopic. θ_i represents the importance of subtopic c_i; $\sum_i \theta_i = 1$. $\mathcal{I}$ is the indicator function. $MaxHeight$ is the height limitation mentioned above. $\Gamma = \gamma^{nrel(c_i, j-1)}$ is the discount factor for novelty, where $nrel(c_i, j-1)$ is the number of relevant documents for subtopic c_i in the previously examined documents (d_1 to d_{j-1}). $Time(D)$ in Formula 1 is the time spent to examine document set D.

The Cube Test [3] is a search effectiveness metric that measures the speed of gaining relevant information (which could be documents or passages) in a dynamic search process. In Cube Test, the user information need regarding a search task is quantified as the volume of a task cube. The cube is compartmentalized into several cells. Each cell represents the information need for a subtopic under the original search task. The size of each cell's bottom area indicates the importance of that subtopic. Finding relevant information for the search task is analogous to 'pouring relevance water into a task cube'. Finding a relevant document helps increase the "relevance water" in the subtopic cells covered by this document. Document relevance is discounted by novelty too – finding a relevant document talking about a subtopic which has been covered by earlier retrieved documents will contribute less relevance water. The height of the cube limits the overall information need for each subtopic and for the overall search task. When the water in a subtopic cell or in the task cube is full, more documents talking about the same subtopic or the same search topic are considered redundant, and will contribute zero relevance. The CubeTest score evaluates the speed of filling up this task cube.

In TREC 2015 DD, we set the novelty parameter $\gamma = 0.5$ and $MaxHeight = 5$. We assume that all subtopics are equally important, and the bottom areas

of each subtopic sum to 1. The time cost $Time(D)$ is counted as the number of interaction iterations between the search engine and the jig for document set D. It means that if a search engine ends the search session earlier, it decreases the user's effort of examining the results and is more likely to gain a higher CT score. If the evaluation stops at the t^{th} iteration, it is denoted as CT@t.

Other metrics used in TREC 2015 DD Track include:

- (P@R) Precision at recall [1] calculates the precision of all results up to the point where every subtopic for a topic is accounted for. $P@R = \frac{r}{R}$, where R is the set of relevant documents in the ground truth file for a search topic and r is number of relevant documents in the top $|R|$ retrieved documents.

- (ERRA) ERR Arithmetic calculates the expected reciprocal rank [2] for each subtopic, and then averages the scores across subtopics using an arithmetic average. $ERR = \sum_{r=1}^{n} \frac{1}{r} \prod_{i=1}^{r-1} (1 - R_i) R_r$, which calculates the probability that the user is not satisfied with the first $r-1$ results and is satisfied with the r^{th} one. It uses a graded relevance for computing stopping probabilities. $R_i = \mathcal{R}(g_i) = \frac{2^{g_i}-1}{2^{g_{max}}}$, where g_i is the relevance grade for the i^{th} document, and g_{max} is the maximum relevance grade for all the relevant documents for a topic.

- (ERRH) ERR Harmonic calculates the expected reciprocal rank for each subtopic, and then averages the scores across subtopics using a harmonic average. It uses graded relevance for computing stopping probabilities.

Note that after the evaluation metrics were released the first time, we have continued to fix bugs and patch the code. The latest metric scripts can be found at `https://github.com/trec-dd/trec-dd-metrics`.

6 Submissions

We have received 32 submissions in total from 7 groups. The groups are listed in Table 3. Below are the system descriptions for the runs. These descriptions were provided by the participating groups at the time that they submitted their runs, and are intended to serve as a roadmap to the proceedings papers from each group.

- baseline BUPT_PRIS main automatic "For each domain, we build an index with INDRI respectively. Given a query, our system runs INDRI once and returns the top 1000 documents in order."

- multir BUPT_PRIS main automatic "To begin, our system returns INDRI results in order. From the 'on_topic' feedback info, our system extracts another query, and runs INDRI again to obtain more documents which are likely to be 'on_topic'. Besides, our system utilizes an TFIDF model to remove documents similar to 'off_topic' ones."

Table 3: Submissions

Group	Country
Beijing University of Posts and Telecommunications (Pattern Recognition and Intelligent System Lab)	China
Georgetown University	USA
Konan University	Japan
Laval University	Canada
University of Glasgow (Terrier Team)	UK
Tianjin University (Institute of Computational Intelligence and Internet)	China
Group Yr	USA

- simti BUPT_PRIS main automatic "Based on the thousands of documents INDRI returns, we build an TFIDF model so as to calculate the similarity between each two documents. To begin, our system returns INDRI results in order. With the feedback info, our system returns documents similar to 'on_topic' ones."

- simtir BUPT_PRIS main automatic "Based on the thousands of documents INDRI returns, we build an TFIDF model so as to calculate the similarity between each two documents. To begin, our system returns INDRI results in order. With the feedback info, our system returns documents similar to 'on_topic' ones and removes documents similar to 'off_topic' ones."

- simtir20 BUPT_PRIS main automatic "Based on the thousands of documents INDRI returns, we build an TFIDF model so as to calculate the similarity between each two documents. To begin, our system returns INDRI results in order. With the feedback info, our system returns documents similar to 'on_topic' ones and removes documents similar to 'off_topic' ones. But before utilizing the info, the top 20 INDRI results must be returned."

- BASE_INDRI_50 DDTJU main automatic "naive version for dd, use Indri index the dataset. for each query, iteration 10 times, each time give the JIGs 5 results"

- ul_lda_roc.2 LavalIVA main automatic "Search engine Solr Algorithms LDA to find subtopics, Roccio to search in other areas and Named Entities recognition in the feedback to reformulate the query."

- ul_lda_roc.3 LavalIVA main automatic "This run uses Solr as a basis for the search engine. We use LDA to search different subtopics in top documents by Solr. And we process the feedback by using a NER algorithm to expand the queries. If there is no feedback we an inverse rocchio algorithm to search in a different area. We used 3 pages of results for this run."

- ul_combi_roc.2 LavalIVA main automatic "For this run we used a combination of results from Solr, LDA and Kmeans with a weight system for each algorithm to contribute to the final top 5 docs to return for the turn. We used an inversed Rocchio algorithm to search in different areas if there is no relevant document in the first turn."

- uogTrEpsilonG uogTr main automatic "S1 (Ranking) Each iteration mixes documents from all indices (weighted by CORI resource ranking). System becomes

less risk averse (will try more documents from low scored resources) if we don't find a relevant document quickly. S2 (Learning) A None S3 (Stopping Condition) First found "

- uogTrRR uogTr main automatic "S1 (Ranking) Round robin for each iteration (5 documents from each domain), where domain ordering is via CORI S2 (Learning) None S3 (Stopping Condition) First found "
- uogTrSI uogTr main automatic "S1 (Ranking) Single index search, each iteration moves down the ranking S2 (Learning) None S3 (Stopping Condition) First found "
- uogTrIL uogTr main automatic "S1 (Ranking) Each iteration mixes documents from all indices, treats each domain evenly. S2 (Learning) None S3 (Stopping Condition) First found "
- ul_lda_roc.10 LavalIVA main automatic "This run uses Solr as a baseline for the search engine. We use LDA to search different subtopics in top documents returned by Solr. And we process the feedback by using a NER algorithm to expand the queries. If there is no feedback we use an inverse rocchio algorithm to search in a different area. We used 10 pages max of results to explore more documents.."
- GU_RUN3_SIMI georgetown main automatic "Lemur+Re-ranking based on document similarity to the feedbacks"
- GU_RUN4_SIMI georgetown main automatic "Lemur+Re-ranking based on document similarity to the feedbacks and the topic"
- tfidf KonanU main automatic "baseline tfidf"
- okapi KonanU main automatic "baseline okapi"
- lm KonanU main automatic "query language model"
- lmrf KonanU main automatic "query language model + relevance feedback"
- uogTrxQuADRR uogTr main automatic "S1 (Ranking) Ranking with xQuAD where pseudo-relevance feedback from the top results is used to generate potential intents. Round robin for each iteration (5 documents from each domain), where domain ordering is via CORI is used S2 (Learning) None S3 (Stopping Condition) First found "
- DDTJU_EXPLORE DDTJU main automatic "explore with the JIGs, more exploitation, less exploration"
- subsimti BUPT_PRIS judged automatic "Based on the thousands of documents INDRI returns, we build an TFIDF model so as to calculate the similarity between each two documents. To begin, our system returns INDRI results in order. With the feedback info, our system returns documents similar to 'on_topic' ones."

- subsimtir BUPT_PRIS judged automatic "Based on the thousands of documents INDRI returns, we build an TFIDF model so as to calculate the similarity between each two documents. To begin, our system returns INDRI results in order. With the feedback info, our system returns documents similar to 'on_topic' ones and removes documents similar to 'off_topic' ones."

- subsimtir20 BUPT_PRIS judged automatic "Based on the thousands of documents INDRI returns, we build an TFIDF model so as to calculate the similarity between each two documents. To begin, our system returns INDRI results in order. With the feedback info, our system returns documents similar to 'on_topic' ones and removes documents similar to 'off_topic' ones. But before utilizing the info, the top 20 INDRI results must be returned."

- subbaseline BUPT_PRIS judged automatic "For each domain, we build an index with INDRI respectively. Given a query, our system runs INDRI once and returns the top 1000 documents in order."

- submultir BUPT_PRIS judged automatic "To begin, our system returns INDRI results in order. From the 'on_topic' feedback info, our system extracts another query, and runs INDRI again to obtain more documents which are likely to be 'on_topic'. Besides, our system utilizes an TFIDF model to remove documents similar to 'off_topic' ones."

- yr_run_with_nov yr judged automatic "With zoning, novelty, and tf-idf"

- yr_run_no_nov yr judged automatic "Zoning, tf-idf, no novelty."

- yr_mixed_sim_nov yr judged automatic "Zoned, confidence averaged between tf-idf and novelty"

- yr_mixed_long yr judged automatic "Zoned, confidence averaged between novelty and similarity. Long runs on each topic in attempt to achieve recall on challenging topics."

- ul_combi_roc_judged LavalIVA judged automatic "Use a combination of results from solr and diversification with lda and kmeans. Use of name entity recognition on feedback and use of inversed rocchio algorithm when there is no feedback."

In summary, many teams used Indri, Lucene, or Terrier as the baseline search systems. Although the Track does not provide subsequent queries in the interactions, extracting new queries from the feedback messages is used in many teams to retrieve documents for the subsequent iterations. Clustering and topic models are used to identify subtopics from the feedback messages, which help with diversification and/or search in specific subtopics in the subsequent retrievals. Teams use heuristics, which might be based on useful user models, to decide when to stop the searches. One challenge that does not seem to have been addressed is varying amounts of relevant information across subtopics.

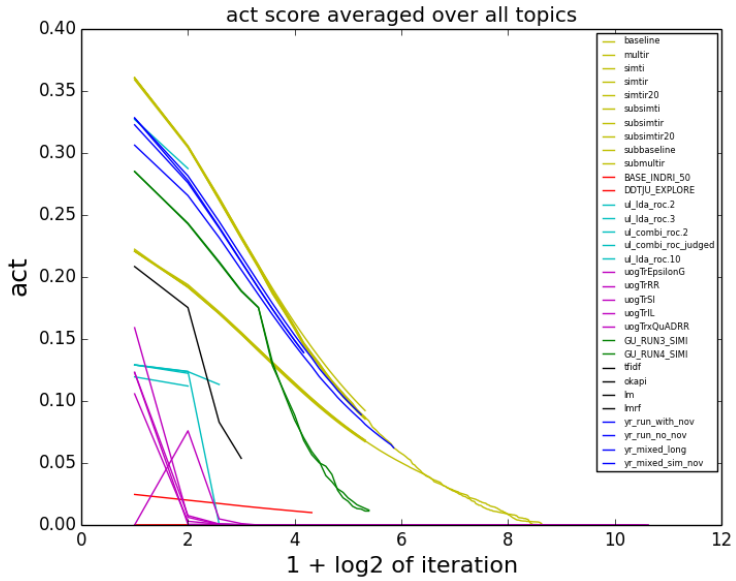

Figure 3: Average Cube Test (ACT) over the Iterations (averaged for all topics).

7 Results

7.1 How Does a System Progress?

Figures 3 and 4 plot the ACT and CT scores for all the runs against the number of search iterations in the dynamic search process. The plots are averaged over all the 118 topics. The plots show a few things. First, they show how a run progresses as the number of iterations increases. In general all runs ACT and CT scores drop, which may indicate that later retrievals are harder so that the speed of getting relevant documents are harder. It may also suggest that how to best make use of the feedback becomes harder as the iterations develop. Noise could be introduced by extracting queries from the feedback messages. Not like in Web search, the user provides fresh queries constantly in a session, here the search engine needs to learn how to get user's intent from the feedback passages, which could be limited if the search engine did not find any relevant documents at the earlier runs. Second, the systems use heuristics and criteria to mark the stopping points. From this year's graphs, we could not see where a good point to stop is partly because the relevance feedback from the first iteration was not sufficiently well used, thus the ACT and CT scores just drop after the first iteration. It looks like a challenge would be how to maintain a constant or even increasing speed of getting relevant documents.

We further plot the gain function of CT, CT, and ACT over the iterations averaged over all topics and all runs in Figure 5. Looking at Figure 5, the

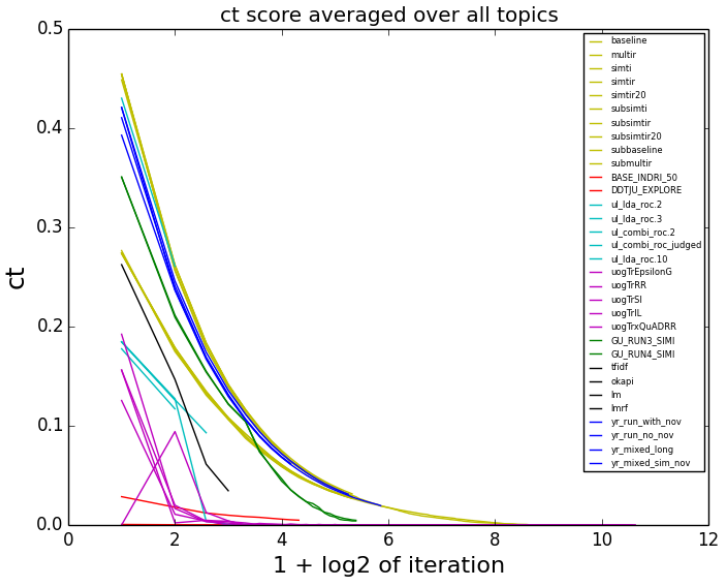

Figure 4: Cube Test (CT) over the Iterations (averaged for all topics).

optimal iteration appears to occur at iteration 10 (ln(iteration) 2.3) as that is where the maximum occurs for Cube Test's Gain function. However, we notice that many runs stop around iteration 10. Therefore, that might be the factor which causes the dropping in the averaged values over all runs.

We observe that ACT is slightly different from CT in that it might be a more stable metric; conceptually it is the average cube-filling speed over every document being retrieved in the entire session. Thus, a system that is looking for a metric that is less sensitive to variation can look at ACT and apply it in the same manner as CT. The concavity of ACT's metric curve would make it easy to identify a good stopping point for the ongoing search iterations.

7.2 Results by Topics

Figures 6 and 7 show the ACT and CT scores for all the topics at the tenth search iteration. At the 10^{th} iteration, most ACT scores are within the range of 0.2 to 0.4; while most CT scores are within the range of less than 0.1. Nonetheless, they both are quite consistent in indicating the topic difficulties.

We also plot all the ACT and CT curves for all runs on every topic. Due to the space limitation of this document, we only show ACT curves for two topics on which most teams score high and two other topics appear to be difficult to many teams. They are shown in Figures 8 to 11.

Topics 15 and 36 are both about searching for named entities, which could be easier for the systems. Topics 62 and 83 are on more general topics, which

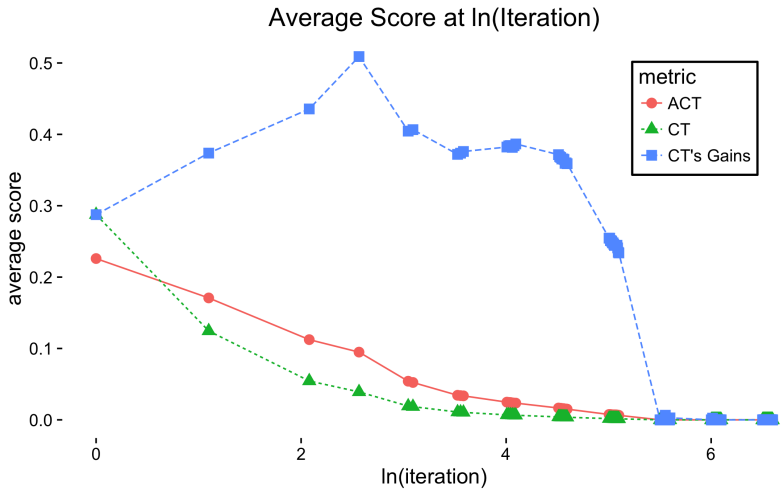

Figure 5: CT, ACT and CT's Gain over the Iterations. Averaged by all runs and all topics.

are more challenging for the systems. We observe that the systems are able to achieve high Cube Test scores at the beginning of a search process on the easy topics and then the speed of getting more relevant documents dramatically drops. On the difficult topics, the curves might be able to climb up when the iterations develop.

7.3 Results by Teams

The evaluation scores for the runs for their first to tenth iterations are listed in Tables 4 to 13. The numbers are averaged over all the topics.

8 Conclusions

The TREC 2015 Dynamic Domain (DD) Track aims to provide evaluation protocols for the complex and dynamic information seeking process. The interactive search process was created to represent the real-world requirements of task-oriented search. We emphasize the entire course of the search process and study the dynamic adaptation of the systems' algorithms to the feedback provided. In the Track's first year, we have gained some initial insights into the interactions between the simulated user and the search systems. In addition to the systems created by track participants, the Track contributes useful training data and tools for annotation, assessing, and scoring. It is quite a challenge to create the assessments at the same time as creating the topics and before the runs were submitted. We hope these tools and data will influence thinking

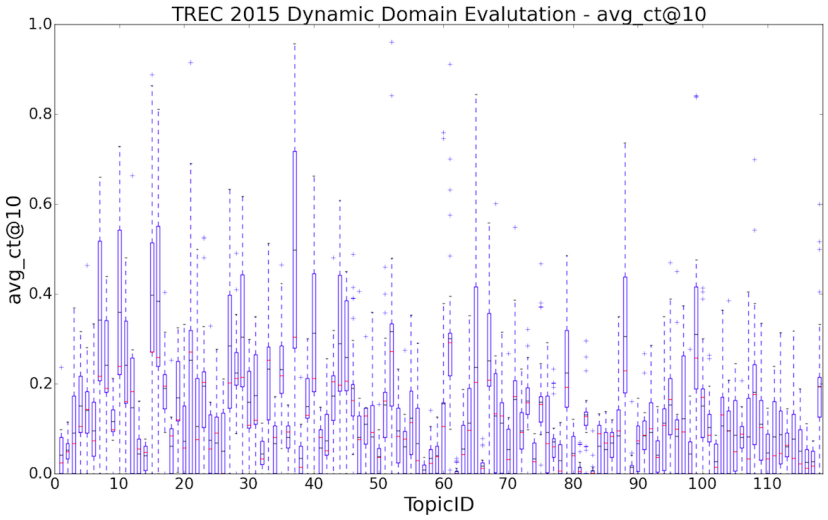

Figure 6: Average Cube Test (ACT) Score at the tenth iteration (By Topics).

in many other activities, such as the knowledge discovery efforts throughout education, government and industry.

9 Acknowledgment

The TREC 2015 Dynamic Domain Track is sponsored by the DARPA Memex program. We thank the following contributors to TREC DD Track in crawling the data: Diffeo, Giant Oak, Hyperion Gray, NASA JPL, and New York University. We have our special thanks to Shiqi Liu for developing the annotation tool, Jiyun Luo for developing the evaluation scripts, Kevin Tian for plotting the graphs. We also thank the joint development efforts from the TREC 2015 Total Recall Track.

References

[1] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *SIGIR '00*.

[2] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM'09*.

[3] J. Luo, C. Wing, H. Yang, and M. Hearst. The water filling model and the cube test: multi-dimensional evaluation for professional search. In *CIKM'13*.

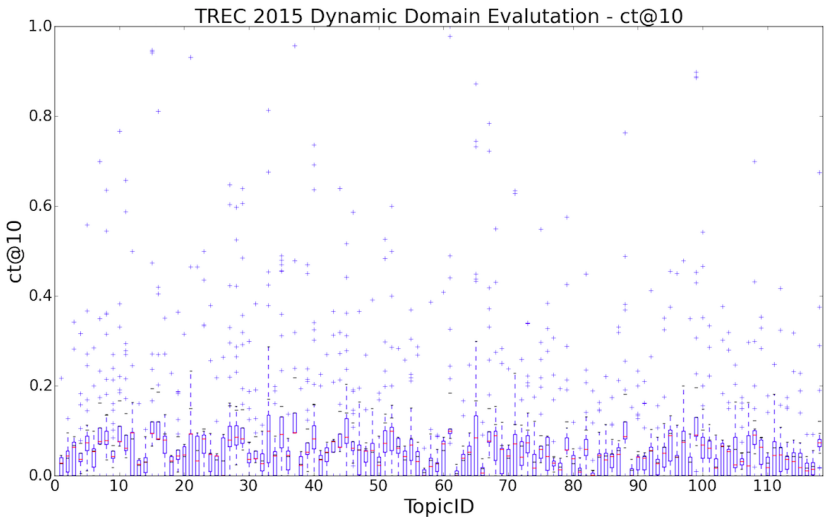

Figure 7: Cube Test (CT) Score at the tenth iteration (By Topics).

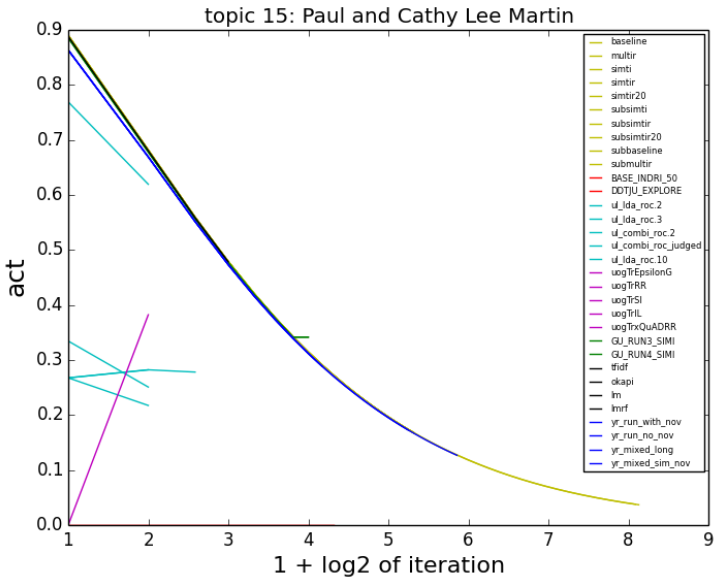

Figure 8: ACT for an easy topic (Topic 15)

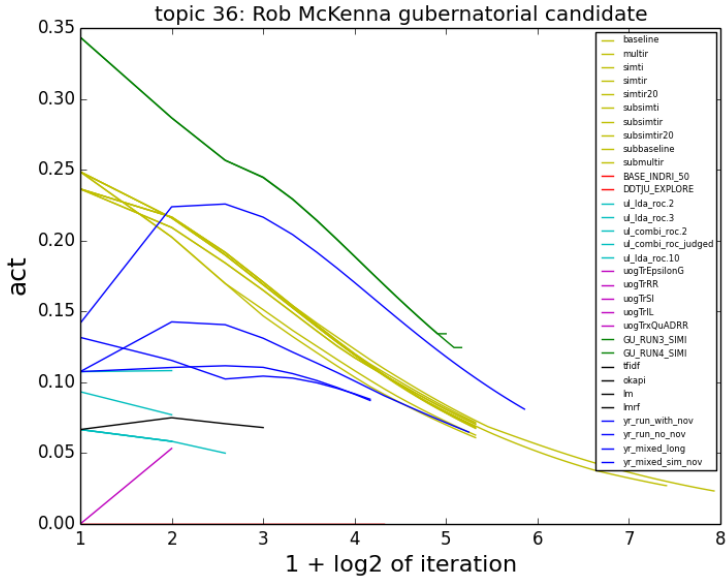

Figure 9: ACT for an easy topic (Topic 36)

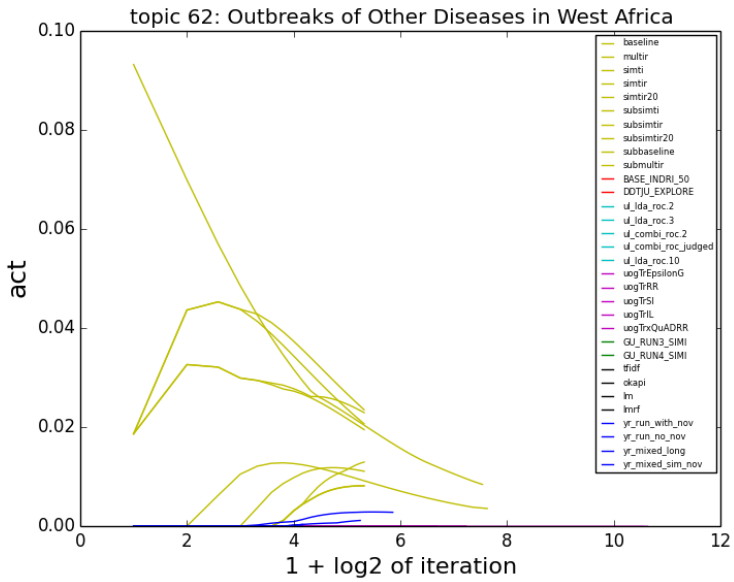

Figure 10: ACT for a difficult topic (Topic 62)

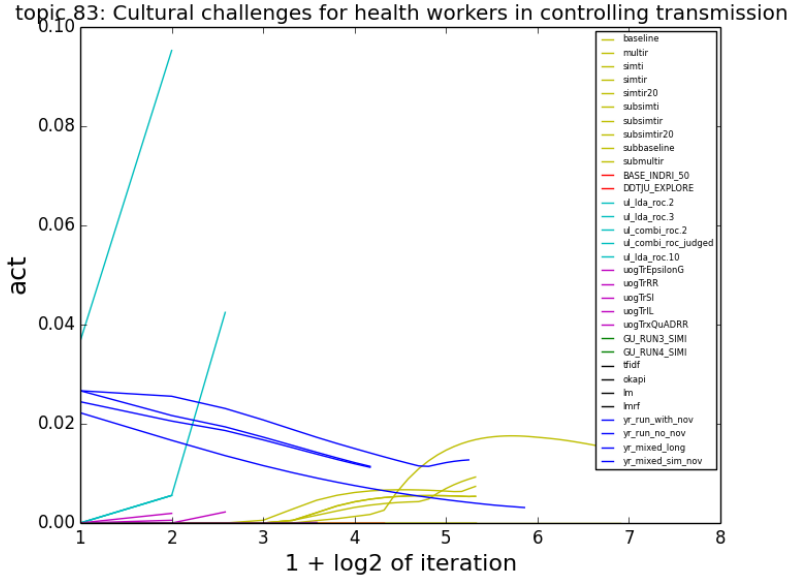

Figure 11: ACT for a difficult topic (Topic 83)

Table 4: TREC 2015 Dynamic Domain Track Evaluation Results - Iteration 1

Run ID	ACT	CT	ERRA	ERRH	P@R
subsimti	0.3662	0.4616	0.4570	0.3750	0.7200
subsimtir20	0.3662	0.4616	0.4570	0.3750	0.7200
subsimtir	0.3662	0.4616	0.4570	0.3750	0.7200
submultir	0.3662	0.4616	0.4570	0.3750	0.7200
subbaseline	0.3636	0.4536	0.4600	0.3810	0.7100
yr_run_with_nov	0.3331	0.4268	0.4660	0.3820	0.7530
yr_run_no_nov	0.3331	0.4268	0.4660	0.3820	0.7530
ul_combi_roc_judged	0.3311	0.4360	0.4240	0.3400	0.6850
yr_mixed_sim_nov	0.3276	0.4165	0.4510	0.3620	0.7440
yr_mixed_long	0.3112	0.3988	0.4380	0.3520	0.7140
baseline	0.2253	0.2802	0.2780	0.2500	0.2320
multir	0.2251	0.2799	0.2870	0.2580	0.2440
simtir	0.2251	0.2799	0.2870	0.2580	0.2440
simtir20	0.2251	0.2799	0.2870	0.2580	0.2440
simti	0.2251	0.2799	0.2870	0.2580	0.2440
lmrf	0.2120	0.2660	0.2970	0.2730	0.2340
lm	0.2115	0.2651	0.2970	0.2730	0.2340
tfidf	0.1876	0.2375	0.2550	0.2270	0.2190
okapi	0.1650	0.2110	0.2310	0.2040	0.2120
uogTrSI	0.1627	0.1959	0.2140	0.1930	0.2050
ul_lda_roc.10	0.1313	0.1882	0.1770	0.1590	0.0860
ul_lda_roc.2	0.1313	0.1882	0.1770	0.1590	0.0860
ul_lda_roc.3	0.1313	0.1882	0.1770	0.1590	0.0860
uogTrRR	0.1259	0.1598	0.1560	0.1390	0.1170
uogTrEpsilonG	0.1259	0.1598	0.1560	0.1390	0.1170
ul_combi_roc.2	0.1231	0.1813	0.1910	0.1760	0.0930
uogTrIL	0.1065	0.1291	0.1490	0.1330	0.0680
DDTJU_EXPLORE	0.0246	0.0286	0.0360	0.0350	0.0250
BASE_INDRI_50	0.0001	0.0006	0.0000	0.0000	0.0000
GU_RUN3_SIMI	0.0000	0.0000	0.0000	0.0000	0.0000
GU_RUN4_SIMI	0.0000	0.0000	0.0000	0.0000	0.0000
uogTrxQuADRR	0.0000	0.0000	0.0000	0.0000	0.0000

Table 5: TREC 2015 Dynamic Domain Track Evaluation Results - Iteration 2

Run ID	ACT	CT	ERRA	ERRH	P@R
subsimtir20	0.3102	0.2656	0.4330	0.3380	0.6470
submultir	0.3102	0.2656	0.4330	0.3380	0.6470
subsimtir	0.3098	0.2609	0.4360	0.3400	0.6510
subsimti	0.3098	0.2609	0.4360	0.3400	0.6510
subbaseline	0.3082	0.2648	0.4350	0.3410	0.6400
ul_combi_roc_judged	0.2914	0.2648	0.3950	0.2890	0.6640
yr_run_no_nov	0.2854	0.2490	0.4330	0.3210	0.6950
yr_run_with_nov	0.2818	0.2401	0.4360	0.3250	0.6410
yr_mixed_sim_nov	0.2799	0.2440	0.4260	0.3200	0.6350
yr_mixed_long	0.2696	0.2397	0.3890	0.2770	0.6090
lm	0.2115	0.2651	0.2970	0.2730	0.2340
simtir	0.1978	0.1826	0.2950	0.2500	0.2070
simti	0.1966	0.1800	0.2940	0.2500	0.2050
multir	0.1956	0.1821	0.2860	0.2420	0.1900
simtir20	0.1956	0.1821	0.2860	0.2420	0.1900
baseline	0.1946	0.1784	0.2810	0.2420	0.1810
tfidf	0.1876	0.2375	0.2550	0.2270	0.2190
lmrf	0.1780	0.1486	0.2910	0.2550	0.1790
uogTrSI	0.1689	0.1146	0.2270	0.2050	0.2050
okapi	0.1650	0.2110	0.2310	0.2040	0.2120
GU_RUN3_SIMI	0.1443	0.1772	0.3490	0.3110	0.3240
GU_RUN4_SIMI	0.1443	0.1772	0.3490	0.3110	0.3240
uogTrRR	0.1336	0.0998	0.1660	0.1480	0.1170
uogTrEpsilonG	0.1265	0.0820	0.1580	0.1420	0.1170
ul_lda_roc.10	0.1261	0.1297	0.1740	0.1500	0.0920
ul_lda_roc.3	0.1258	0.1288	0.1740	0.1500	0.0920
ul_lda_roc.2	0.1246	0.1285	0.1720	0.1460	0.0860
ul_combi_roc.2	0.1148	0.1189	0.1900	0.1640	0.0800
uogTrIL	0.1095	0.0754	0.1540	0.1370	0.0680
uogTrxQuADRR	0.0786	0.0981	0.2110	0.1950	0.1610
DDTJU_EXPLORE	0.0200	0.0179	0.0390	0.0380	0.0130
BASE_INDRI_50	0.0002	0.0003	0.0000	0.0000	0.0000

Table 6: TREC 2015 Dynamic Domain Track Evaluation Results - Iteration 3

Run ID	ACT	CT	ERRA	ERRH	P@R
ul_combi_roc_judged	0.2914	0.2648	0.3950	0.2890	0.6640
subsimtir20	0.2678	0.1859	0.4200	0.3100	0.6050
submultir	0.2678	0.1859	0.4200	0.3100	0.6050
subsimtir	0.2662	0.1819	0.4320	0.3270	0.6080
subbaseline	0.2661	0.1850	0.4230	0.3140	0.5990
subsimti	0.2661	0.1816	0.4320	0.3260	0.6030
yr_run_no_nov	0.2480	0.1768	0.4050	0.2710	0.6470
yr_run_with_nov	0.2436	0.1713	0.4200	0.2940	0.5870
yr_mixed_sim_nov	0.2426	0.1711	0.4100	0.2870	0.5810
yr_mixed_long	0.2351	0.1691	0.3790	0.2530	0.5340
lm	0.2115	0.2651	0.2970	0.2730	0.2340
tfidf	0.1876	0.2375	0.2550	0.2270	0.2190
simtir	0.1754	0.1347	0.2910	0.2380	0.1780
multir	0.1741	0.1371	0.2790	0.2270	0.1640
simtir20	0.1741	0.1371	0.2790	0.2270	0.1640
simti	0.1739	0.1333	0.2900	0.2370	0.1750
baseline	0.1730	0.1361	0.2730	0.2240	0.1580
uogTrSI	0.1694	0.0796	0.2280	0.2070	0.2050
lmrf	0.1665	0.1025	0.2860	0.2450	0.1660
okapi	0.1650	0.2110	0.2310	0.2040	0.2120
GU_RUN4_SIMI	0.1402	0.1426	0.3410	0.2870	0.2700
GU_RUN3_SIMI	0.1398	0.1412	0.3440	0.2960	0.2690
uogTrRR	0.1344	0.0704	0.1680	0.1500	0.1170
uogTrEpsilonG	0.1272	0.0595	0.1610	0.1440	0.1170
ul_lda_roc.10	0.1267	0.0904	0.1760	0.1520	0.0920
ul_lda_roc.2	0.1246	0.1285	0.1720	0.1460	0.0860
ul_lda_roc.3	0.1153	0.0946	0.1750	0.1510	0.0890
ul_combi_roc.2	0.1148	0.1189	0.1900	0.1640	0.0800
uogTrIL	0.1102	0.0542	0.1570	0.1400	0.0680
uogTrxQuADRR	0.0835	0.0780	0.2190	0.2030	0.1610
DDTJU_EXPLORE	0.0174	0.0121	0.0390	0.0380	0.0080
BASE_INDRI_50	0.0002	0.0003	0.0000	0.0000	0.0000

Table 7: TREC 2015 Dynamic Domain Track Evaluation Results - Iteration 4

Run ID	ACT	CT	ERRA	ERRH	P@R
ul_combi_roc_judged	0.2914	0.2648	0.3950	0.2890	0.6640
subsimtir20	0.2362	0.1431	0.4140	0.2920	0.5670
submultir	0.2362	0.1431	0.4140	0.2920	0.5670
subbaseline	0.2349	0.1436	0.4140	0.2920	0.5610
subsimtir	0.2347	0.1415	0.4220	0.3070	0.5550
subsimti	0.2345	0.1411	0.4230	0.3070	0.5510
yr_run_no_nov	0.2201	0.1379	0.3940	0.2460	0.6040
yr_run_with_nov	0.2155	0.1328	0.4080	0.2750	0.5400
yr_mixed_sim_nov	0.2149	0.1338	0.3980	0.2610	0.5350
lm	0.2115	0.2651	0.2970	0.2730	0.2340
yr_mixed_long	0.2086	0.1310	0.3730	0.2430	0.4860
tfidf	0.1876	0.2375	0.2550	0.2270	0.2190
uogTrSI	0.1697	0.0637	0.2300	0.2090	0.2050
okapi	0.1650	0.2110	0.2310	0.2040	0.2120
lmrf	0.1600	0.0776	0.2850	0.2460	0.1550
simtir	0.1584	0.1102	0.2850	0.2220	0.1520
multir	0.1575	0.1098	0.2780	0.2250	0.1440
simtir20	0.1575	0.1098	0.2780	0.2250	0.1440
simti	0.1567	0.1083	0.2830	0.2200	0.1500
baseline	0.1563	0.1081	0.2720	0.2180	0.1380
uogTrRR	0.1344	0.0528	0.1680	0.1500	0.1170
GU_RUN4_SIMI	0.1314	0.1175	0.3300	0.2580	0.2290
GU_RUN3_SIMI	0.1308	0.1168	0.3310	0.2610	0.2300
uogTrEpsilonG	0.1274	0.0480	0.1630	0.1470	0.1170
ul_lda_roc.10	0.1268	0.0686	0.1770	0.1520	0.0920
ul_lda_roc.2	0.1246	0.1285	0.1720	0.1460	0.0860
ul_lda_roc.3	0.1153	0.0946	0.1750	0.1510	0.0890
ul_combi_roc.2	0.1148	0.1189	0.1900	0.1640	0.0800
uogTrIL	0.1106	0.0426	0.1580	0.1410	0.0680
uogTrxQuADRR	0.0847	0.0627	0.2220	0.2050	0.1610
DDTJU_EXPLORE	0.0154	0.0098	0.0370	0.0350	0.0060
BASE_INDRI_50	0.0002	0.0002	0.0000	0.0000	0.0000

Table 8: TREC 2015 Dynamic Domain Track Evaluation Results - Iteration 5

Run ID	ACT	CT	ERRA	ERRH	P@R
ul_combi_roc_judged	0.2914	0.2648	0.3950	0.2890	0.6640
subsimtir20	0.2126	0.1183	0.4100	0.2850	0.5410
submultir	0.2123	0.1177	0.4100	0.2860	0.5280
lm	0.2115	0.2651	0.2970	0.2730	0.2340
subbaseline	0.2114	0.1173	0.4120	0.2920	0.5280
subsimtir	0.2109	0.1150	0.4190	0.3010	0.5170
subsimti	0.2106	0.1143	0.4220	0.3060	0.5100
yr_run_no_nov	0.1985	0.1134	0.3870	0.2410	0.5690
yr_run_with_nov	0.1938	0.1077	0.4040	0.2710	0.4980
yr_mixed_sim_nov	0.1936	0.1087	0.3950	0.2570	0.4960
yr_mixed_long	0.1882	0.1073	0.3690	0.2330	0.4550
tfidf	0.1876	0.2375	0.2550	0.2270	0.2190
uogTrSI	0.1697	0.0510	0.2300	0.2090	0.2050
okapi	0.1650	0.2110	0.2310	0.2040	0.2120
lmrf	0.1600	0.0776	0.2850	0.2460	0.1550
simtir	0.1447	0.0905	0.2820	0.2160	0.1340
simtir20	0.1445	0.0934	0.2750	0.2070	0.1410
multir	0.1444	0.0933	0.2730	0.2040	0.1400
simti	0.1430	0.0888	0.2820	0.2170	0.1300
baseline	0.1428	0.0908	0.2710	0.2190	0.1220
uogTrRR	0.1345	0.0431	0.1690	0.1510	0.1170
uogTrEpsilonG	0.1276	0.0405	0.1640	0.1470	0.1170
ul_lda_roc.10	0.1269	0.0556	0.1770	0.1530	0.0920
ul_lda_roc.2	0.1246	0.1285	0.1720	0.1460	0.0860
GU_RUN4_SIMI	0.1227	0.0983	0.3260	0.2480	0.2020
GU_RUN3_SIMI	0.1222	0.0988	0.3230	0.2440	0.2050
ul_lda_roc.3	0.1153	0.0946	0.1750	0.1510	0.0890
ul_combi_roc.2	0.1148	0.1189	0.1900	0.1640	0.0800
uogTrIL	0.1106	0.0342	0.1580	0.1410	0.0680
uogTrxQuADRR	0.0847	0.0502	0.2220	0.2050	0.1610
DDTJU_EXPLORE	0.0140	0.0085	0.0360	0.0310	0.0060
BASE_INDRI_50	0.0002	0.0002	0.0000	0.0000	0.0000

Table 9: TREC 2015 Dynamic Domain Track Evaluation Results - Iteration 6

Run ID	ACT	CT	ERRA	ERRH	P@R
ul_combi_roc_judged	0.2914	0.2648	0.3950	0.2890	0.6640
lm	0.2115	0.2651	0.2970	0.2730	0.2340
subsimtir20	0.1940	0.0999	0.4070	0.2800	0.5100
submultir	0.1934	0.0999	0.4040	0.2750	0.4900
subbaseline	0.1929	0.0994	0.4050	0.2760	0.4930
subsimtir	0.1921	0.0972	0.4120	0.2850	0.4830
subsimti	0.1918	0.0971	0.4140	0.2900	0.4760
tfidf	0.1876	0.2375	0.2550	0.2270	0.2190
yr_run_no_nov	0.1812	0.0957	0.3850	0.2410	0.5370
yr_run_with_nov	0.1766	0.0911	0.3970	0.2590	0.4680
yr_mixed_sim_nov	0.1765	0.0915	0.3940	0.2570	0.4640
yr_mixed_long	0.1718	0.0903	0.3680	0.2300	0.4290
uogTrSI	0.1697	0.0431	0.2300	0.2090	0.2050
okapi	0.1650	0.2110	0.2310	0.2040	0.2120
lmrf	0.1600	0.0776	0.2850	0.2460	0.1550
uogTrRR	0.1345	0.0359	0.1690	0.1510	0.1170
simtir20	0.1337	0.0803	0.2740	0.2040	0.1290
multir	0.1336	0.0804	0.2710	0.2030	0.1290
simtir	0.1334	0.0773	0.2770	0.2050	0.1190
baseline	0.1319	0.0784	0.2710	0.2160	0.1120
simti	0.1318	0.0766	0.2750	0.2050	0.1190
uogTrEpsilonG	0.1276	0.0341	0.1640	0.1480	0.1170
ul_lda_roc.10	0.1269	0.0464	0.1770	0.1530	0.0920
ul_lda_roc.2	0.1246	0.1285	0.1720	0.1460	0.0860
GU_RUN4_SIMI	0.1170	0.0850	0.3220	0.2450	0.1860
GU_RUN3_SIMI	0.1167	0.0847	0.3220	0.2430	0.1870
ul_lda_roc.3	0.1153	0.0946	0.1750	0.1510	0.0890
ul_combi_roc.2	0.1148	0.1189	0.1900	0.1640	0.0800
uogTrIL	0.1106	0.0289	0.1580	0.1410	0.0680
uogTrxQuADRR	0.0849	0.0434	0.2240	0.2070	0.1610
DDTJU_EXPLORE	0.0129	0.0077	0.0370	0.0320	0.0050
BASE_INDRI_50	0.0002	0.0001	0.0000	0.0000	0.0000

Table 10: TREC 2015 Dynamic Domain Track Evaluation Results - Iteration 7

Run ID	ACT	CT	ERRA	ERRH	P@R
ul_combi_roc_judged	0.2914	0.2648	0.3950	0.2890	0.6640
lm	0.2115	0.2651	0.2970	0.2730	0.2340
tfidf	0.1876	0.2375	0.2550	0.2270	0.2190
subsimtir20	0.1788	0.0860	0.4070	0.2790	0.4750
submultir	0.1781	0.0860	0.4040	0.2730	0.4590
subbaseline	0.1778	0.0859	0.4040	0.2730	0.4690
subsimtir	0.1769	0.0844	0.4060	0.2720	0.4490
subsimti	0.1766	0.0841	0.4110	0.2850	0.4430
uogTrSI	0.1698	0.0375	0.2310	0.2090	0.2050
yr_run_no_nov	0.1671	0.0825	0.3850	0.2410	0.5120
okapi	0.1650	0.2110	0.2310	0.2040	0.2120
yr_run_with_nov	0.1626	0.0788	0.3930	0.2510	0.4440
yr_mixed_sim_nov	0.1626	0.0792	0.3880	0.2450	0.4390
lmrf	0.1600	0.0776	0.2850	0.2460	0.1550
yr_mixed_long	0.1584	0.0781	0.3670	0.2270	0.4100
uogTrRR	0.1345	0.0311	0.1690	0.1510	0.1170
uogTrEpsilonG	0.1277	0.0296	0.1650	0.1480	0.1170
ul_lda_roc.10	0.1269	0.0400	0.1780	0.1530	0.0920
ul_lda_roc.2	0.1246	0.1285	0.1720	0.1460	0.0860
simtir20	0.1245	0.0697	0.2730	0.2010	0.1170
multir	0.1245	0.0700	0.2710	0.2030	0.1160
simtir	0.1239	0.0676	0.2740	0.2000	0.1070
baseline	0.1228	0.0685	0.2670	0.2040	0.1050
simti	0.1226	0.0674	0.2740	0.2020	0.1060
ul_lda_roc.3	0.1153	0.0946	0.1750	0.1510	0.0890
ul_combi_roc.2	0.1148	0.1189	0.1900	0.1640	0.0800
GU_RUN4_SIMI	0.1127	0.0735	0.3220	0.2450	0.1710
GU_RUN3_SIMI	0.1123	0.0733	0.3220	0.2430	0.1700
uogTrIL	0.1106	0.0249	0.1580	0.1410	0.0680
uogTrxQuADRR	0.0849	0.0382	0.2240	0.2070	0.1610
DDTJU_EXPLORE	0.0120	0.0067	0.0370	0.0320	0.0050
BASE_INDRI_50	0.0002	0.0002	0.0000	0.0000	0.0000

Table 11: TREC 2015 Dynamic Domain Track Evaluation Results - Iteration 8

Run ID	ACT	CT	ERRA	ERRH	P@R
ul_combi_roc_judged	0.2914	0.2648	0.3950	0.2890	0.6640
lm	0.2115	0.2651	0.2970	0.2730	0.2340
tfidf	0.1876	0.2375	0.2550	0.2270	0.2190
uogTrSI	0.1698	0.0336	0.2310	0.2100	0.2050
subsimtir20	0.1663	0.0761	0.4020	0.2690	0.4500
subbaseline	0.1654	0.0758	0.4020	0.2710	0.4420
submultir	0.1653	0.0761	0.3990	0.2610	0.4320
okapi	0.1650	0.2110	0.2310	0.2040	0.2120
subsimtir	0.1645	0.0746	0.4020	0.2670	0.4230
subsimti	0.1642	0.0748	0.4060	0.2770	0.4150
lmrf	0.1600	0.0776	0.2850	0.2460	0.1550
yr_run_no_nov	0.1553	0.0729	0.3830	0.2370	0.4930
yr_mixed_sim_nov	0.1510	0.0698	0.3850	0.2390	0.4200
yr_run_with_nov	0.1509	0.0695	0.3880	0.2420	0.4250
yr_mixed_long	0.1472	0.0690	0.3650	0.2230	0.3940
uogTrRR	0.1345	0.0272	0.1690	0.1510	0.1170
uogTrEpsilonG	0.1277	0.0262	0.1650	0.1480	0.1170
ul_lda_roc.10	0.1269	0.0353	0.1780	0.1530	0.0920
ul_lda_roc.2	0.1246	0.1285	0.1720	0.1460	0.0860
simtir20	0.1166	0.0619	0.2710	0.1990	0.1080
multir	0.1166	0.0617	0.2710	0.1990	0.1060
simtir	0.1159	0.0600	0.2730	0.1970	0.0990
ul_lda_roc.3	0.1153	0.0946	0.1750	0.1510	0.0890
baseline	0.1151	0.0612	0.2660	0.2030	0.0960
ul_combi_roc.2	0.1148	0.1189	0.1900	0.1640	0.0800
simti	0.1147	0.0597	0.2730	0.1980	0.0960
uogTrIL	0.1106	0.0219	0.1590	0.1410	0.0680
GU_RUN4_SIMI	0.1093	0.0649	0.3200	0.2420	0.1620
GU_RUN3_SIMI	0.1088	0.0649	0.3200	0.2400	0.1620
uogTrxQuADRR	0.0850	0.0337	0.2240	0.2070	0.1610
DDTJU_EXPLORE	0.0112	0.0059	0.0370	0.0320	0.0040
BASE_INDRI_50	0.0002	0.0002	0.0000	0.0000	0.0000

Table 12: TREC 2015 Dynamic Domain Track Evaluation Results - Iteration 9

Run ID	ACT	CT	ERRA	ERRH	P@R
ul_combi_roc_judged	0.2914	0.2648	0.3950	0.2890	0.6640
lm	0.2115	0.2651	0.2970	0.2730	0.2340
tfidf	0.1876	0.2375	0.2550	0.2270	0.2190
uogTrSI	0.1699	0.0299	0.2320	0.2100	0.2050
okapi	0.1650	0.2110	0.2310	0.2040	0.2120
lmrf	0.1600	0.0776	0.2850	0.2460	0.1550
subsimtir20	0.1558	0.0680	0.4010	0.2680	0.4230
subbaseline	0.1550	0.0678	0.4010	0.2690	0.4190
submultir	0.1544	0.0680	0.3980	0.2580	0.4080
subsimtir	0.1541	0.0669	0.4000	0.2660	0.4020
subsimti	0.1538	0.0670	0.4030	0.2690	0.3930
yr_run_no_nov	0.1453	0.0654	0.3810	0.0460	0.4700
yr_mixed_sim_nov	0.1411	0.0627	0.3810	0.2330	0.4010
yr_run_with_nov	0.1411	0.0622	0.3870	0.2400	0.4040
yr_mixed_long	0.1377	0.0617	0.3640	0.2240	0.3780
uogTrRR	0.1345	0.0249	0.1690	0.1510	0.1170
uogTrEpsilonG	0.1277	0.0237	0.1650	0.1480	0.1170
ul_lda_roc.10	0.1269	0.0314	0.1780	0.1530	0.0920
ul_lda_roc.2	0.1246	0.1285	0.1720	0.1460	0.0860
ul_lda_roc.3	0.1153	0.0946	0.1750	0.1510	0.0890
ul_combi_roc.2	0.1148	0.1189	0.1900	0.1640	0.0800
uogTrIL	0.1107	0.0202	0.1590	0.1420	0.0680
simtir20	0.1098	0.0556	0.2700	0.1980	0.0990
multir	0.1097	0.0550	0.2710	0.1990	0.0970
simtir	0.1090	0.0539	0.2720	0.1950	0.0910
baseline	0.1084	0.0549	0.2650	0.2010	0.0900
simti	0.1079	0.0537	0.2710	0.1950	0.0890
GU_RUN4_SIMI	0.1066	0.0580	0.3200	0.2420	0.1560
GU_RUN3_SIMI	0.1063	0.0580	0.3190	0.2400	0.1560
uogTrxQuADRR	0.0850	0.0300	0.2250	0.2080	0.1610
DDTJU_EXPLORE	0.0106	0.0053	0.0370	0.0320	0.0040
BASE_INDRI_50	0.0002	0.0002	0.0000	0.0000	0.0000

Table 13: TREC 2015 Dynamic Domain Track Evaluation Results - Iteration 10

Run ID	ACT	CT	ERRA	ERRH	P@R
ul_combi_roc_judged	0.2914	0.2648	0.3950	0.2890	0.6640
lm	0.2115	0.2651	0.2970	0.2730	0.2340
tfidf	0.1876	0.2375	0.2550	0.2270	0.2190
uogTrSI	0.1699	0.0269	0.2320	0.2100	0.2050
okapi	0.1650	0.2110	0.2310	0.2040	0.2120
lmrf	0.1600	0.0776	0.2850	0.2460	0.1550
subsimtir20	0.1469	0.0615	0.3990	0.2680	0.3980
subbaseline	0.1461	0.0619	0.4010	0.2690	0.3980
subsimtir	0.1453	0.0607	0.3970	0.2660	0.3790
submultir	0.1451	0.0615	0.3980	0.2580	0.3860
subsimti	0.1450	0.0606	0.4010	0.2670	0.3730
yr_mixed_sim_nov	0.1411	0.0627	0.3810	0.2330	0.4010
yr_run_with_nov	0.1411	0.0622	0.3870	0.2400	0.4040
yr_run_no_nov	0.1367	0.0592	0.3790	0.2310	0.4500
uogTrRR	0.1346	0.0229	0.1690	0.1520	0.1170
yr_mixed_long	0.1295	0.0558	0.3630	0.2230	0.3620
uogTrEpsilonG	0.1277	0.0215	0.1650	0.1480	0.1170
ul_lda_roc.10	0.1269	0.0283	0.1780	0.1530	0.0920
ul_lda_roc.2	0.1246	0.1285	0.1720	0.1460	0.0860
ul_lda_roc.3	0.1153	0.0946	0.1750	0.1510	0.0890
ul_combi_roc.2	0.1148	0.1189	0.1900	0.1640	0.0800
uogTrIL	0.1107	0.0184	0.1590	0.1420	0.0680
GU_RUN4_SIMI	0.1045	0.0525	0.3200	0.2420	0.1490
GU_RUN3_SIMI	0.1043	0.0532	0.3180	0.2380	0.1520
simtir20	0.1039	0.0506	0.2700	0.1970	0.0920
multir	0.1038	0.0502	0.2700	0.1980	0.0900
simtir	0.1030	0.0493	0.2710	0.1950	0.0850
baseline	0.1026	0.0506	0.2620	0.1890	0.0850
simti	0.1019	0.0487	0.2710	0.1950	0.0830
uogTrxQuADRR	0.0850	0.0272	0.2250	0.2080	0.1610
DDTJU_EXPLORE	0.0100	0.0049	0.0350	0.0280	0.0040
BASE_INDRI_50	0.0002	0.0002	0.0000	0.0000	0.0000