TREC 2017 Dynamic Domain Track Overview

Grace Hui Yang
Georgetown University
huiyang@cs.georgetown.edu

Zhiwen Tang
Georgetown University
zt79@georgetown.edu

Ian Soboroff
NIST
ian.soboroff@nist.gov

1. Introduction

The goal of dynamic domain track is promoting the research of dynamic, exploratory search within complex information domains, where the search process is usually interactive and user's information need is also complex. Dynamic Domain (DD) track has been held in the past three years. This track's name includes two parts. "Dynamic" means the search process may contain multiple runs of iteration, and the participating system is expected to adapt its search algorithm based on the relevance feedback. "Domain" means the search task focuses on special domains, where user's information need consists of multiple aspects, and the participating system is expected to help the user explore the domain through rich interaction. This task has received great attention and this track is inspired by interested groups in government, including DARPA MEMEX program.

The settings of DD track are motivated by professional search, such as prior art patent search or criminal network finding, where rich interaction is a great asset for improving the search results and users require stringent relevancy on the documents returned.

In order to simulate the interaction between the search engine and the user, as well as evaluate the whole search process, a simulated user (called Jig[1]) is developed. During each step in interaction, participating system sends a list of documents to the simulated user, and the simulated user returns a real-time feedback to the participating system. Participating system learns the real intention behind the search topic, adapts its search algorithm and generates the next list of documents. This process is repeated until the participating system believes the user's information need has been satisfied. All the documents returned by the participating system are saved for the evaluation of the whole search session.

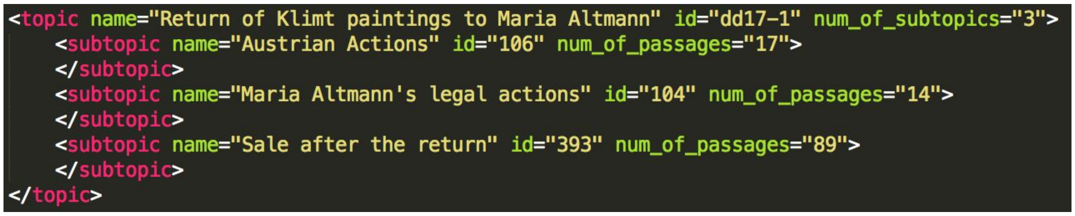
DD track also uses fine-grained judgements. Different from open domain web search, all the relevance judgements are on the passage level, which expresses user's information need more accurately. Correspondingly, DD track also uses sophisticated metrics to evaluate the search results, which includes Cube Test [1], session-DCG [2] and Expected Utility [3]. All these metrics evaluate the whole search process and each provides a distinct view on the effectiveness and efficiency of the participating systems.

This year, DD track focuses on the exploration of New York Times archives [4]. 3 groups participated and 11 runs were submitted.

2. Task Description

The task of TREC DD track is based on a concept that search is driven by the feedback instead of queries. That is, a good search system need to learn user's real intents through the feedback given by the user regarding previous returned documents and the user does not need to reformulate queries to express or refine his/her information need. The search system needs to adjust its search algorithm so as to help the user explore the complex domain.

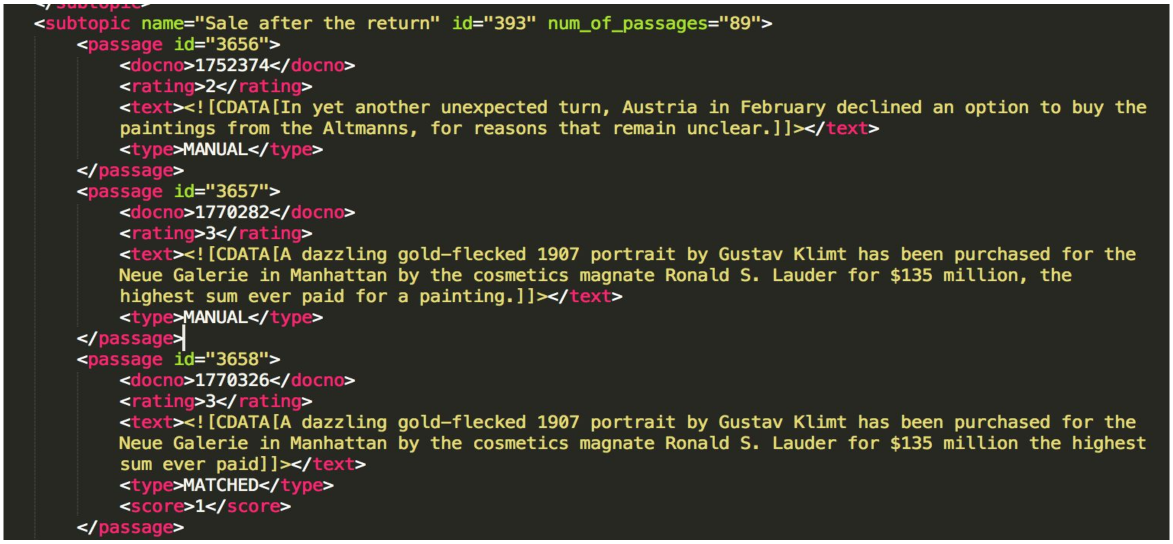
[1] https://github.com/trec-dd/trec-dd-jig



```
<topic name="Return of Klimt paintings to Maria Altmann" id="dd17-1" num_of_subtopics="3">
    <subtopic name="Austrian Actions" id="106" num_of_passages="17">
    </subtopic>
    <subtopic name="Maria Altmann's legal actions" id="104" num_of_passages="14">
    </subtopic>
    <subtopic name="Sale after the return" id="393" num_of_passages="89">
    </subtopic>
</topic>
```

Figure 1. Sample topic of TREC 2017 DD track

Every search topic in DD track contains multiple aspects, which are referred as subtopics in this track. All the relevance judgements are on passage level. Every document may contain several relevant passages, each related to a subtopic with different relevance scores.

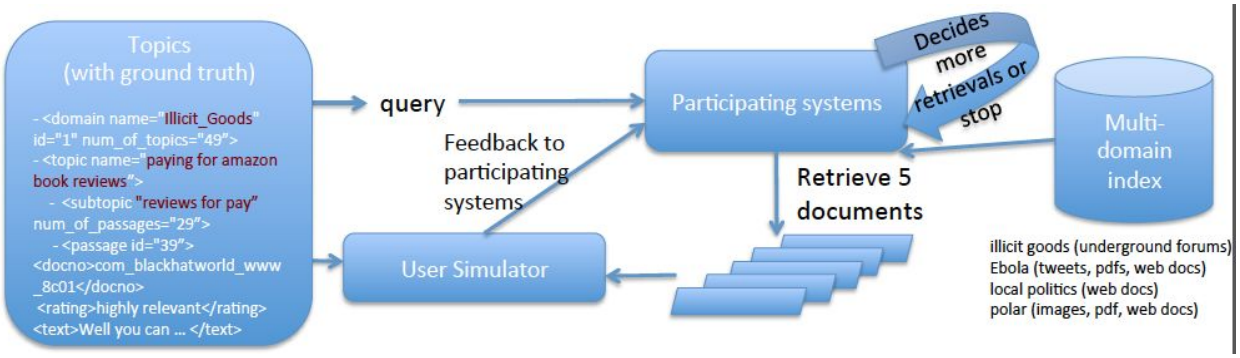


```
    </subtopic>
    <subtopic name="Sale after the return" id="393" num_of_passages="89">
        <passage id="3656">
            <docno>1752374</docno>
            <rating>2</rating>
            <text><![CDATA[In yet another unexpected turn, Austria in February declined an option to buy the
            paintings from the Altmanns, for reasons that remain unclear.]]></text>
            <type>MANUAL</type>
        </passage>
        <passage id="3657">
            <docno>1770282</docno>
            <rating>3</rating>
            <text><![CDATA[A dazzling gold-flecked 1907 portrait by Gustav Klimt has been purchased for the
            Neue Galerie in Manhattan by the cosmetics magnate Ronald S. Lauder for $135 million, the
            highest sum ever paid for a painting.]]></text>
            <type>MANUAL</type>
        </passage>
        <passage id="3658">
            <docno>1770326</docno>
            <rating>3</rating>
            <text><![CDATA[A dazzling gold-flecked 1907 portrait by Gustav Klimt has been purchased for the
            Neue Galerie in Manhattan by the cosmetics magnate Ronald S. Lauder for $135 million the highest
            sum ever paid]]></text>
            <type>MATCHED</type>
            <score>1</score>
        </passage>
```

Figure 2. Sample Relevance Judgement of TREC 2017 DD track

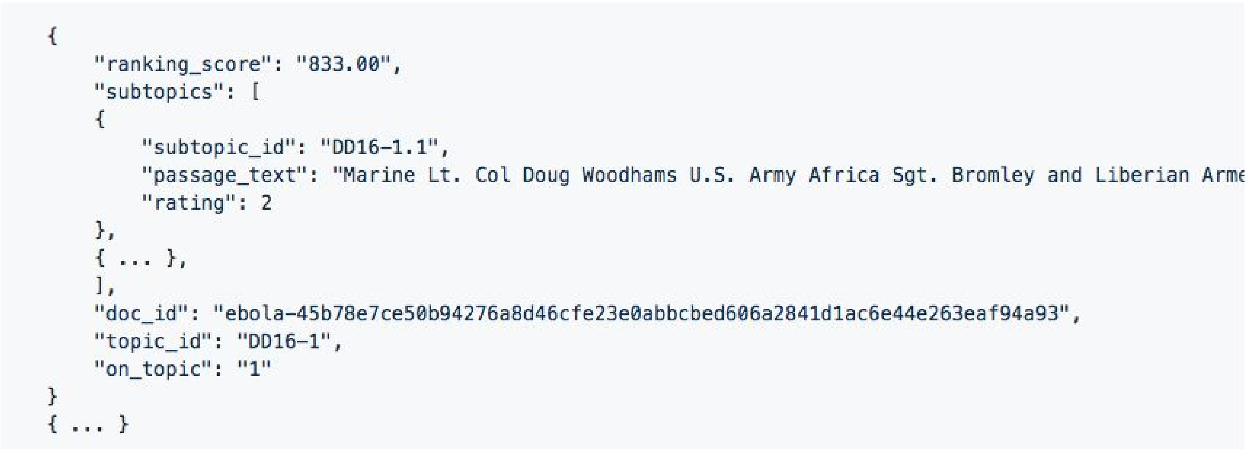
In the beginning, the search system receives an initial query (the topic name) indicating user's intention. Then, the search system retrieves five documents from the index and sends them back to the user simulator Jig. Jig returns feedback about the returned five documents. The feedback sent back by Jig gives a detailed description about the relevance of returned documents on subtopic level. The search system needs to decide if it will continue returning documents or stop the current search. The search system may also consider how to rerank the documents so as to better satisfy user's information need.

This track is expecting participating systems achieve two basic points. First, the search system is expected to adapt its search algorithm based on the relevance feedback, for which rich information is provided in the feedback. Second, it is the search system's job, instead of the user's, to decide whether to continue search. Search results are evaluated using different sophisticated metrics, which measures the total amount of information the search system gains and the effort of user from various points of view.



```
- <domain name="Illicit_Goods"
id="1" num_of_topics="49">
- <topic name="paying for amazon
book reviews">
   - <subtopic "reviews for pay"
num_of_passages="29">
   - <passage id="39">
<docno>com_blackhatworld_www
_8c01</docno>
 <rating>highly relevant</rating>
<text>Well you can ... </text>
```

Figure 3. DD track Task illustration



```
{
    "ranking_score": "833.00",
    "subtopics": [
    {
        "subtopic_id": "DD16-1.1",
        "passage_text": "Marine Lt. Col Doug Woodhams U.S. Army Africa Sgt. Bromley and Liberian Arme
        "rating": 2
    },
    { ... },
    ],
    "doc_id": "ebola-45b78e7ce50b94276a8d46cfe23e0abbcbed606a2841d1ac6e44e263eaf94a93",
    "topic_id": "DD16-1",
    "on_topic": "1"
}
{ ... }
```

Figure 4. Sample feedback returned by Jig

3. Dataset and Groundtruth

3.1. Dataset

In 2017, DD track focuses on the exploring of a new domain, the archives of New York Times in 20 years [4]. The corpus contains all the articles published in New York Times (online and offline) from January 1, 1987 to June 19, 2007 with metadata provided by the New York Times Newsroom, the New York Times Indexing Service and the online production staff at nytimes.com. Most articles are manually summarized and tagged by professional staffs. The original form of this dataset is in News Industry Text Format (NITF)[2]. This corpus contains huge amount of information covering a wide range of topics and categories.

Compressed Size	Uncompressed Size	Number of Documents	Number of Queries
3.1 GB	16 GB	1855658	60

Table 1. Statistics of New York Time dataset

[2] https://iptc.org/standards/nitf/

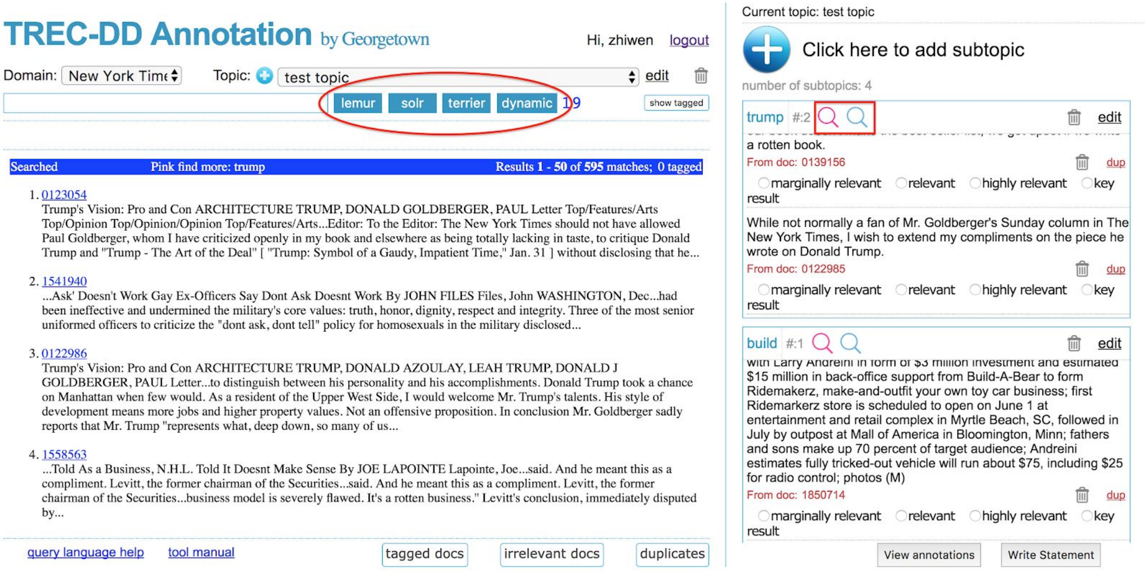
3.2. Topic Development

Topics are developed by six NIST assessors in over six weeks during the summer of 2017. A topic, which is like a query, is the main search target for the whole search process. Every topic contains several subtopics, each addresses one aspect of the search topic. Each subtopic contains several number of passages which is discovered from the entire corpus. Each passage is graded based on the relevance between its content and the subtopic. An annotation tool is developed to help the assessors find the complete set of passages that are relevant to the query. Also, near-duplicate detection is utilized to help find possible relevant passages that may be missed. The graded passages are treated as the complete set of judgement.

The user interface of annotation tool is shown in Figure 5 and Figure 6. Four algorithms are provided for search in topic level and two algorithms are used for search in subtopic level. Assessors first conduct search in topic level. They can then go into the detailed page of every document where they will give rich feedback to the annotation tool. They can decide if a document is irrelevant or duplicated. They can also drag and drop relevant passages to the subtopic box on the right side. Each passage is then graded based on the extent of relevance. The highest relevance score is 4 for key results and the lowest relevance score is 1 for marginally relevant.

In the topic level search, three mainstream open source search engine, Lemur[3], Solr[4] and Terrier[5], and an active learning search algorithm are provided. DD track expects the combination of different search engines can reduce the inherent bias of each individual one. The active learning algorithm uses the feedback, i.e. the documents graded by assessors before, to refine its search results. For every query, assessors are required to search in every one of the four search buttons and go over the ranking lists so as to cover as many relevant documents as possible.

In the subtopic level search, two search algorithms are used to help assessors find more relevant passages. One of them uses the passages that have been tagged and the subtopic name to search for relevant documents, the other one only uses tagged passages to search. For every subtopic, assessors are also required to search in each one of them so as to find a complete set of relevant passages.



1. 0123054
Trump's Vision: Pro and Con ARCHITECTURE TRUMP, DONALD GOLDBERGER, PAUL Letter Top/Features/Arts Top/Opinion Top/Opinion/Opinion Top/Features/Arts...Editor: To the Editor: The New York Times should not have allowed Paul Goldberger, whom I have criticized openly in my book and elsewhere as being totally lacking in taste, to critique Donald Trump and "Trump - The Art of the Deal" ["Trump: Symbol of a Gaudy, Impatient Time," Jan. 31] without disclosing that he...

2. 1541940
...Ask' Doesn't Work Gay Ex-Officers Say Dont Ask Doesnt Work By JOHN FILES Files, John WASHINGTON, Dec...had been ineffective and undermined the military's core values: truth, honor, dignity, respect and integrity. Three of the most senior uniformed officers to criticize the "dont ask, dont tell" policy for homosexuals in the military disclosed...

3. 0122986
Trump's Vision: Pro and Con ARCHITECTURE TRUMP, DONALD AZOULAY, LEAH TRUMP, DONALD J GOLDBERGER, PAUL Letter...to distinguish between his personality and his accomplishments. Donald Trump took a chance on Manhattan when few would. As a resident of the Upper West Side, I would welcome Mr. Trump's talents. His style of development means more jobs and higher property values. Not an offensive proposition. In conclusion Mr. Goldberger sadly reports that Mr. Trump "represents what, deep down, so many of us...

4. 1558563
...Told As a Business, N.H.L. Told It Doesnt Make Sense By JOE LAPOINTE Lapointe, Joe...said. And he meant this as a compliment. Levitt, the former chairman of the Securities...said. And he meant this as a compliment. Levitt, the former chairman of the Securities...business model is severely flawed. It's a rotten business." Levitt's conclusion, immediately disputed by...

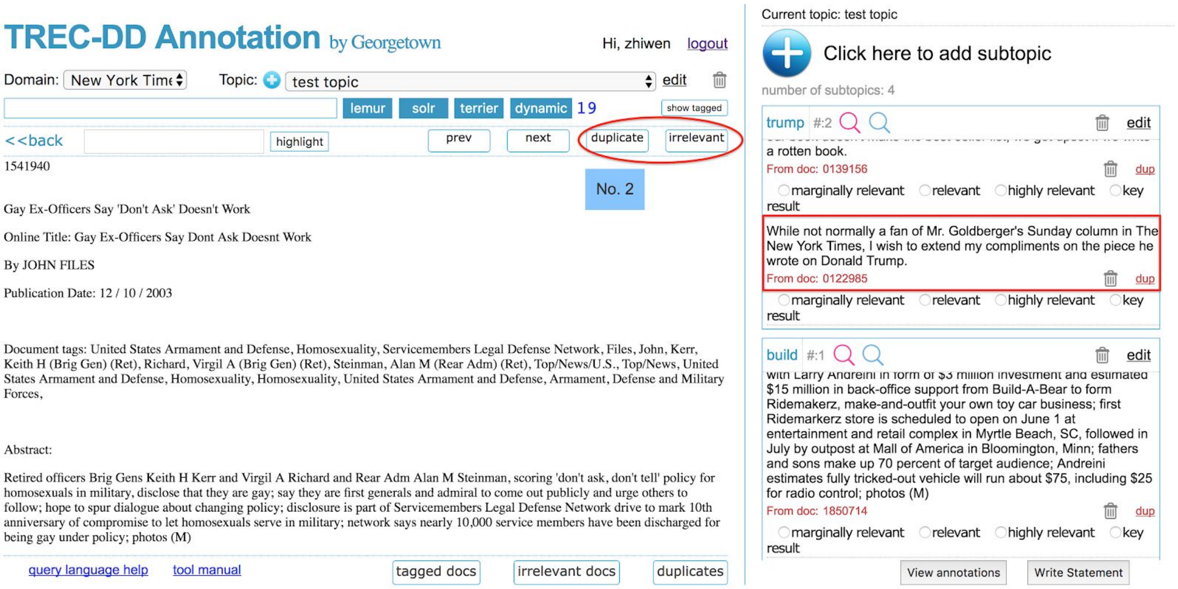
a rotten book.
From doc: 0139156 dup
○marginally relevant ○relevant ○highly relevant ○key result

While not normally a fan of Mr. Goldberger's Sunday column in The New York Times, I wish to extend my compliments on the piece he wrote on Donald Trump.
From doc: 0122985 dup
○marginally relevant ○relevant ○highly relevant ○key result

with Larry Andreini in form of $3 million investment and estimated $15 million in back-office support from Build-A-Bear to form Ridemakerz, make-and-outfit your own toy car business; first Ridemarkerz store is scheduled to open on June 1 at entertainment and retail complex in Myrtle Beach, SC, followed in July by outpost at Mall of America in Bloomington, Minn; fathers and sons make up 70 percent of target audience; Andreini estimates fully tricked-out vehicle will run about $75, including $25 for radio control; photos (M)
From doc: 1850714 dup
○marginally relevant ○relevant ○highly relevant ○key result

Figure 5. User Interface(I) of Annotation Tool

[3] http://www.lemurproject.org/
[4] http://lucene.apache.org/solr/
[5] http://terrier.org/



1541940

Gay Ex-Officers Say 'Don't Ask' Doesn't Work

Online Title: Gay Ex-Officers Say Dont Ask Doesnt Work

By JOHN FILES

Publication Date: 12 / 10 / 2003

Document tags: United States Armament and Defense, Homosexuality, Servicemembers Legal Defense Network, Files, John, Kerr, Keith H (Brig Gen) (Ret), Richard, Virgil A (Brig Gen) (Ret), Steinman, Alan M (Rear Adm) (Ret), Top/News/U.S., Top/News, United States Armament and Defense, Homosexuality, Homosexuality, United States Armament and Defense, Armament, Defense and Military Forces,

Abstract:

Retired officers Brig Gens Keith H Kerr and Virgil A Richard and Rear Adm Alan M Steinman, scoring 'don't ask, don't tell' policy for homosexuals in military, disclose that they are gay; say they are first generals and admiral to come out publicly and urge others to follow; hope to spur dialogue about changing policy; disclosure is part of Servicemembers Legal Defense Network drive to mark 10th anniversary of compromise to let homosexuals serve in military; network says nearly 10,000 service members have been discharged for being gay under policy; photos (M)

a rotten book.
From doc: 0139156

While not normally a fan of Mr. Goldberger's Sunday column in The New York Times, I wish to extend my compliments on the piece he wrote on Donald Trump.
From doc: 0122985

with Larry Andreini in form of $3 million investment and estimated $15 million in back-office support from Build-A-Bear to form Ridemakerz, make-and-outfit your own toy car business; first Ridemarkerz store is scheduled to open on June 1 at entertainment and retail complex in Myrtle Beach, SC, followed in July by outpost at Mall of America in Bloomington, Minn; fathers and sons make up 70 percent of target audience; Andreini estimates fully tricked-out vehicle will run about $75, including $25 for radio control; photos (M)
From doc: 1850714

Figure 6. User Interface(II) of Annotation Tool

After completing the human annotation, near-duplicate detection is used to find passages that are actually relevant but may be missed by assessors. In the ground truth data, passages that are tagged with “MANUAL” are those discovered by human assessors while passages that are tagged with “MATCHED” are those discovered using detection algorithm, which is based on its similarity to the human annotated passages.

4. Evaluation metrics

In order to provide a comprehensive evaluation of the whole search process, in 2017, DD track uses several evaluation metrics to measure the performance of search systems from different perspectives. The primary metric used in DD track is Cube Test [1]. DD track also uses session-DCG [2] and Expected Utility [3] for evaluation. These metrics reveal different aspects of the dynamic search process.

Cube Test (CT) [1] is a search effectiveness measurement evaluating the speed of gaining relevant information (could be documents or passages) in a dynamic search process. It measures the amount of relevant information a system could gather and the time needed in the entire search process. The higher the Cube Test score, the better the IR system. It is defined as:

$$CT = \frac{\sum_{i=1}^{L} \sum_{j=1}^{|list_i|} \sum_{c} \theta_c * rel_c(i,j) * \gamma^{n(c,i,j-1)} * I(\sum_{m=1}^{i-1} \sum_{n=1}^{|list_m|} rel_c(m,n) + \sum_{n=1}^{j-1} rel_c(i,n) < MaxHeight)}{L}$$

where L is the number of iterations so far, $|list_i|$ is the number of documents returned at the i^{th} iteration. c is a subtopic, θ_c is the importance factor of subtopic c, $rel_c(i,j)$ is the relevance score of the j^{th} document that is returned in i^{th} iteration regarding subtopic c, $n(c,i,j-1)$ is the number of relevant documents found on subtopic c before the j^{th} document in the i^{th} iteration, γ is the discounting factor, $I(*)$ is an indicator function and $MaxHeight$ is the cap for each subtopic.

Session-DCG (sDCG) [2] extends the classic DCG to a search session which consists of multiple iterations. The relevance scores of results that are ranked lower or returned in later iterations get more discounts. The discounted cumulative relevance score is the final result of this metric. It is defined as :

$$sDCG = \sum_{i=1}^{L} \sum_{j=1}^{|list_i|} \frac{rel(i,j)}{(1+log_b j) * (1+log_{bq} i)}$$

sDCG does not consider subtopics so $rel(i, j)$ is the relevance score that is accumulated over all the subtopics. Both b and bq are discounting factors.

Expected Utility (EU) [3] scores different runs by measuring the relevant information a system found and the length of documents. The relevance scores of documents are discounted based on ranking order and novelty. The document length is discounted only based on ranking position. The difference between the cumulative relevance score and the aggregated document length is the final score of each run. It is defined as:

$$EU = \sum_{\omega} P(\omega) \left(\sum_{(i,j)\in\omega} \left(\sum_{c\in d_{i,j}} \theta_c * \gamma^{n(c,\,i,\,j-1)} \right) - a * len(i, j) \right)$$

EU assumes that the user only reviews a subset (ω) of documents returned. $P(\omega)$ is the probability of subset ω being reviewed, $len(i,j)$ is the length of document that is ranked j^{th} in the i^{th} iteration and a is the coefficient.

Apart from the raw scores of these metrics, DD track also uses normalized scores of these metrics following the methods proposed in [5] where the upper bound scores of every topic are used for normalization. DD track expects the normalized scores bringing more fairness to the dynamic search evaluation.

The parameters used for DD track evaluation are as follows: In CT, $\gamma = 0.5$ and $MaxHeight = 5$. In sDCG, $b = 2$ and $bq = 4$. In EU, $\gamma = 0.5$ and $a = 0.01$. All the subtopics within the same topic are assumed to be equally important.

5. Submission and Results

5.1. Submission

In 2017, 3 groups participated in the DD track and 11 runs are submitted in total.

Group	Country
University of Maryland (CLIP)	USA
Georgetown University (georgetown)	USA
Chinese Academy of Science (ICTNET)	China

Table 2. Participating Groups

Here are the brief summary of submitted runs provided by the participating groups:

clip_addwords: Data was indexed/searched using Indri search engine. For the first run, the topic was used as the search term. For subsequent runs, words from the passage text (excluding stopwords) were added to the query.

clip_baseline: Baseline set of results using the topic description as the search terms (indexed/searched using Indri search engine). The top 25 results for each topic were submitted to the jig, 5 per topic for each run.

clip_filter: Data was indexed/searched using Indri search engine. For the first run, topics were used as search terms. For subsequent runs, Indri filter operator was used to add terms from the relevant passages (provided via the jig). The topic terms were used as "required" terms, and words from the passages were only valid if the other terms also appeared.

dqn_5_actions: Use DQN to choose 5 possible search actions

dqn_semantic_state: state is defined as query+feedback+iteration number

galago_baseline: The first 50 results returned by galago

ictnet_div_qe: We use xQuAD and query expansion algorithm to ensure both relevance and diversification. Use stop strategy. The first iteration we use the result of solr.

ictnet_emulti: For ebola dataset, we use google suggested queries and jig feedback to ensure the diversification. For New York Times dataset, most of the queries is long and there is no suggested queries, we only use feedback information. We use xQuAD and query expansion algorithms

ictnet_fom_itr1: In this solution, we run xQuAD and query expansion algorithms which is sim with other solutions. But we change parameters.

ictnet_params1_s: Change params of other solutions. Use stop strategy.

ictnet_params2_ns: Change params of other solution. Not use stop strategy.

5.2. Results

The evaluation scores, including the raw scores and normalized scores, in the first ten iterations of all submitted runs are plotted from Figure 7 to Figure 12. More detailed results can be found from Table 3 to Table 12.

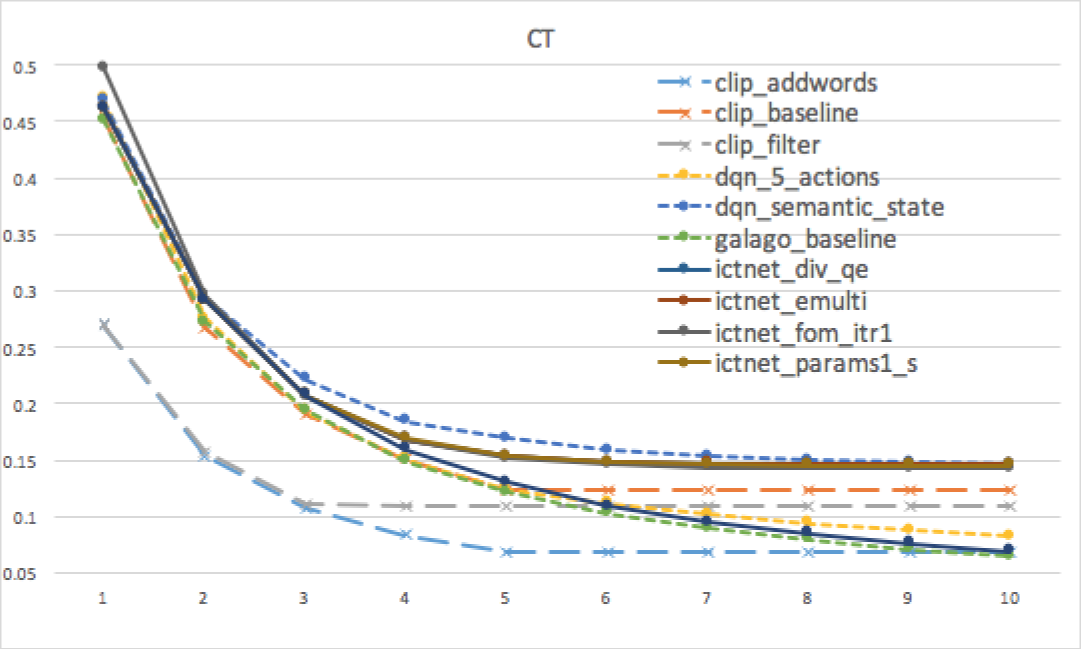


Figure 7. CT scores in the first ten iterations

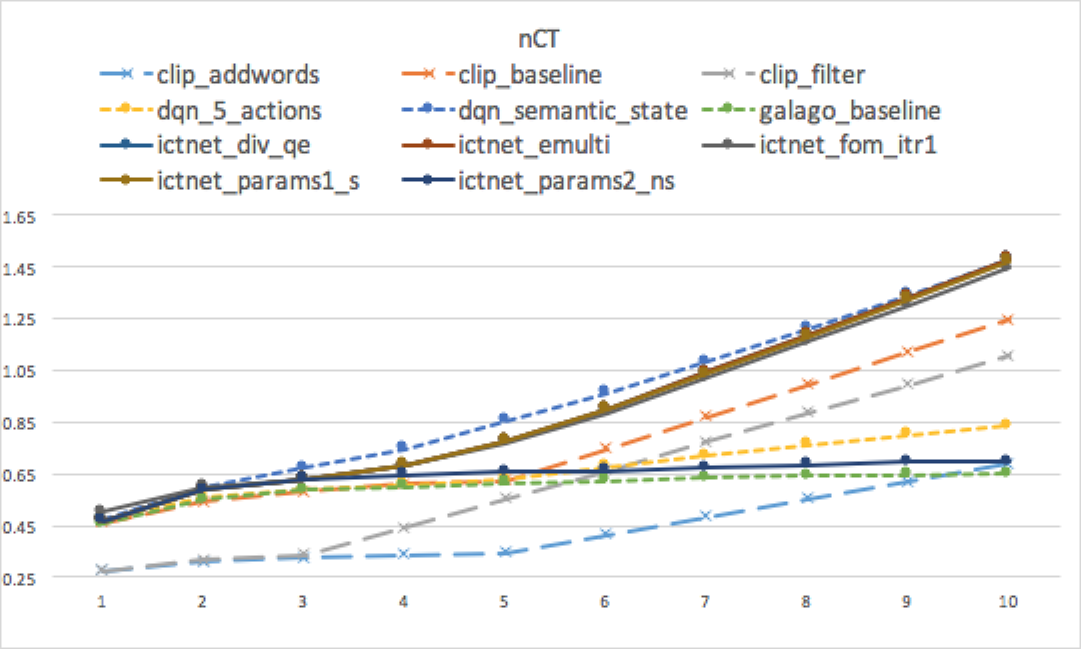


Figure 8. Normalized CT scores in the first ten iterations

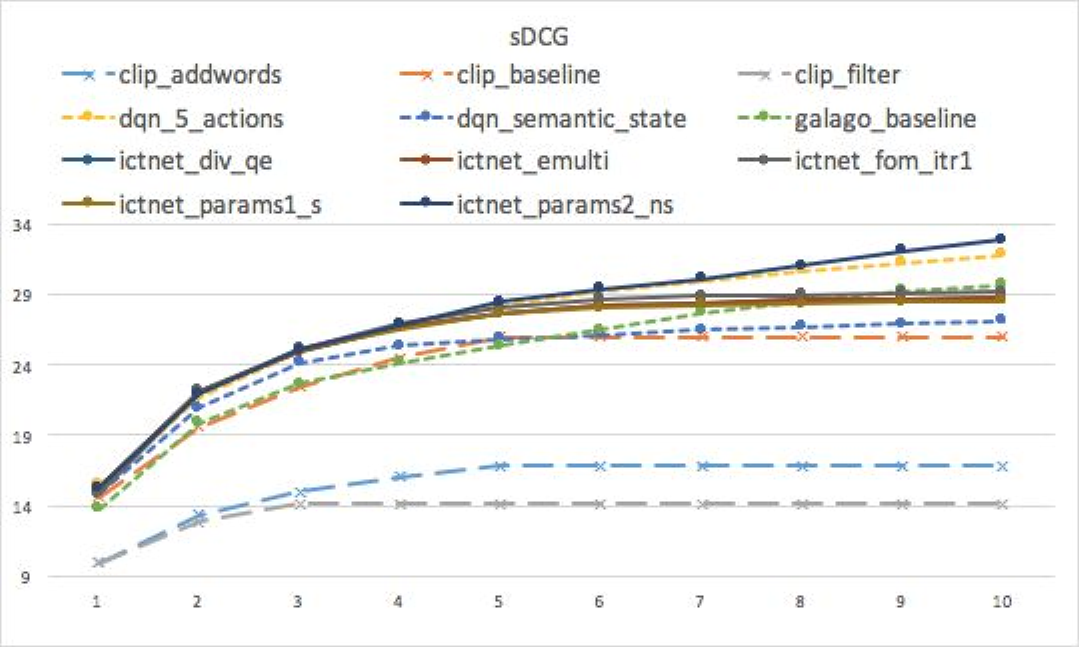


Figure 9. sDCG scores in the first ten iterations

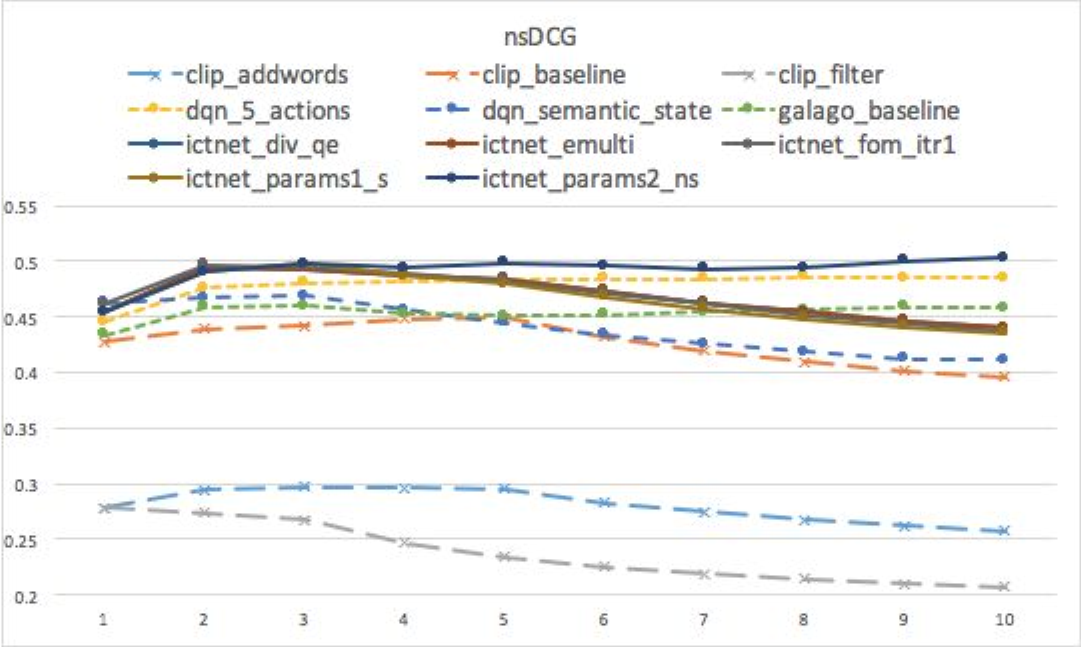


Figure 10. Normalized sDCG scores in the first ten iterations

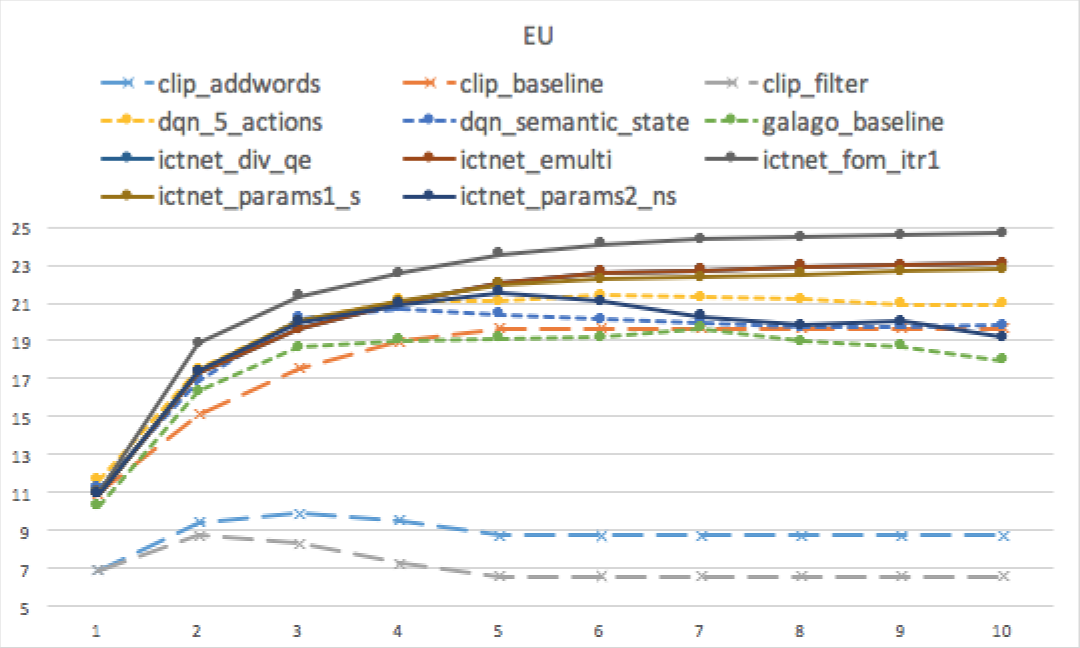


Figure 11. EU scores in the first ten iterations

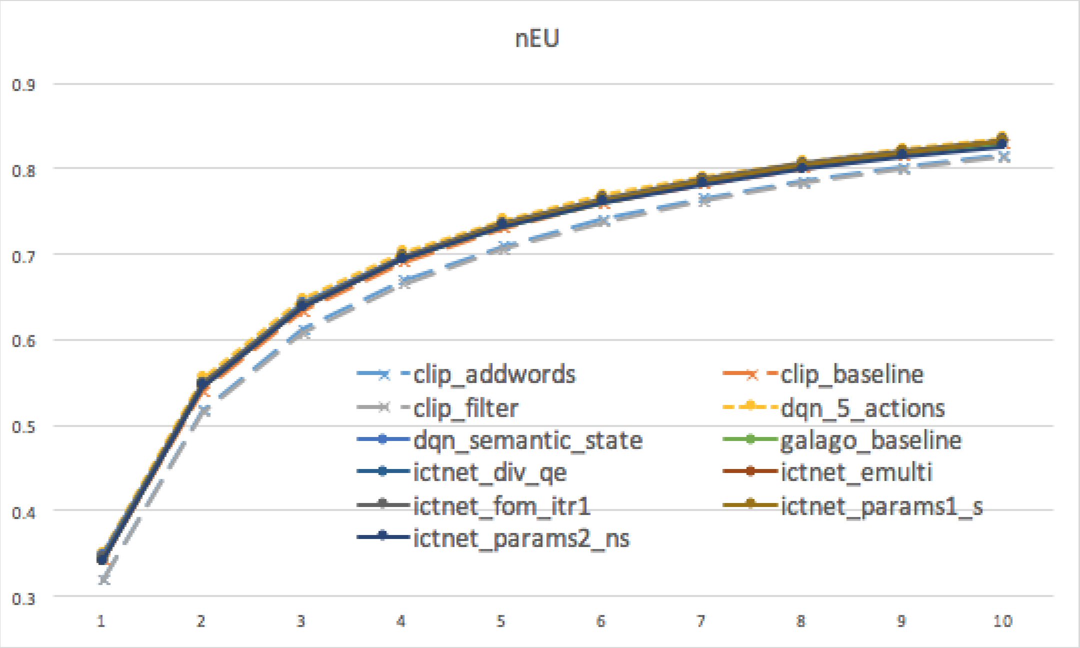


Figure 12. Normalized EU scores in the first ten iterations

6. Discussion

The dynamic domain track has been running in the past three years at TREC and this is the final year. DD track always focuses on interactive and exploratory search task with the vision that a good search system should be a guide to discover the domain of user's interest by adaptively learning user's intention. Participants have tried various methods marching towards this goal.

With the rising of Artificial Intelligence (AI) in recent years, it is exciting to see that this track shares a very similar goal with many AI tasks, that is, building an intelligent system that understands human's mind better. Although DD track ends this year, as organizers, we are still hoping researchers to keep interests in this task, especially the potential improvement brought by the latest AI techniques.

Acknowledgement

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

Reference

[1] Luo, Jiyun, Christopher Wing, Hui Yang, and Marti 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, pp. 709-714. ACM, 2013.

[2] Järvelin, Kalervo, Susan Price, Lois Delcambre, and Marianne Nielsen. "Discounted cumulated gain based evaluation of multiple-query IR sessions." Advances in Information Retrieval (2008): 4-15.

[3] Yang, Yiming, and Abhimanyu Lad. "Modeling expected utility of multi-session information distillation." In Conference on the Theory of Information Retrieval, pp. 164-175. Springer, Berlin, Heidelberg, 2009.

[4] Sandhaus, Evan. "The new york times annotated corpus." Linguistic Data Consortium, Philadelphia 6, no. 12 (2008): e26752.

[5] Tang, Zhiwen, and Grace Hui Yang. "Investigating per Topic Upper Bound for Session Search Evaluation." In Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, pp. 185-192. ACM, 2017.

Detailed Results

	CT	nCT	sDCG	nsDCG	EU	nEU
clip_addwords	0.269316	0.2729377	9.9550701	0.2779112	6.8649358	0.3182756
clip_baseline	0.4537569	0.4591389	14.5994951	0.4274217	10.950894	0.342133
clip_filter	0.269316	0.2729377	9.9550701	0.2779112	6.8649358	0.3182756
dqn_5_actions	0.4701114	0.4756687	15.4846276	0.4456217	11.6124865	0.3499484
dqn_semantic_state	0.4683404	0.4742194	14.8987069	0.4620971	11.1862801	0.3473296
galago_baseline	0.450791	0.4563555	13.7844676	0.4337153	10.248345	0.3431296
ictnet_div_qe	0.4618715	0.4677867	15.2283842	0.454452	10.8843135	0.3407438
ictnet_emulti	0.4618715	0.4677867	15.2283842	0.454452	10.8843135	0.3407438
ictnet_fom_itr1	0.4975938	0.5034492	14.8143864	0.4617947	11.0105412	0.3435733
ictnet_params1_s	0.4618715	0.4677867	15.2283842	0.454452	10.8843135	0.3407438
ictnet_params2_ns	0.4618715	0.4677867	15.2283842	0.454452	10.8843135	0.3407438

Table 3. Iteration 1

	CT	nCT	sDCG	nsDCG	EU	nEU
clip_addwords	0.1531085	0.3102113	13.3407969	0.2942657	9.4279455	0.5176928
clip_baseline	0.2666988	0.5395375	19.5664954	0.4394219	15.1047319	0.5388618
clip_filter	0.1564809	0.3175536	12.8289596	0.2739706	8.7072546	0.5152451
dqn_5_actions	0.2760114	0.5592316	21.6554912	0.4760824	17.4928539	0.5543615
dqn_semantic_state	0.2938375	0.5938134	20.9093702	0.4666493	16.8901066	0.5473597
galago_baseline	0.2722861	0.5507842	19.8558578	0.4589931	16.3226741	0.5476611
ictnet_div_qe	0.2916159	0.5902519	22.0556315	0.4948688	17.3078898	0.5449988
ictnet_emulti	0.2916159	0.5902519	22.0556315	0.4948688	17.3078898	0.5449988
ictnet_fom_itr1	0.2955616	0.5979995	22.1477807	0.4970647	18.8855788	0.5491104
ictnet_params1_s	0.2912779	0.5894399	21.9231463	0.4901891	17.3714137	0.5450564
ictnet_params2_ns	0.2912779	0.5894399	21.9231463	0.4901891	17.3714137	0.5450564

Table 4. Iteration 2

	CT	nCT	sDCG	nsDCG	EU	nEU
clip_addwords	0.1073827	0.3264266	15.0027584	0.2965049	9.8808072	0.6115602
clip_baseline	0.1904297	0.5782188	22.4560498	0.441601	17.5553643	0.6326569
clip_filter	0.1106119	0.337056	14.0906933	0.2668653	8.3000888	0.6071541
dqn_5_actions	0.1930831	0.5867918	24.8716411	0.4804548	20.0101208	0.6468404
dqn_semantic_state	0.2217546	0.6720723	24.1273948	0.4687206	20.1680049	0.6419633
galago_baseline	0.193695	0.5879702	22.6569676	0.4594243	18.6602872	0.639583
ictnet_div_qe	0.2077662	0.6307294	24.9336764	0.4927374	19.6095602	0.6371803
ictnet_emulti	0.2077662	0.6307294	24.9336764	0.4927374	19.6095602	0.6371803
ictnet_fom_itr1	0.2064842	0.6266709	25.1365731	0.4950392	21.3302725	0.6405269
ictnet_params1_s	0.2069918	0.6282322	25.0580261	0.4970876	20.0124248	0.6382257
ictnet_params2_ns	0.2069918	0.6282322	25.0580261	0.4970876	19.9629873	0.638051

Table 5. Iteration 3

	CT	nCT	sDCG	nsDCG	EU	nEU
clip_addwords	0.0831655	0.3371635	16.0390466	0.2961623	9.5226217	0.6696075
clip_baseline	0.1509181	0.6109704	24.4608545	0.4484322	18.9589704	0.6905211
clip_filter	0.108693	0.4417074	14.0906933	0.2468105	7.2210982	0.6644384
dqn_5_actions	0.1501568	0.6087354	26.8278014	0.4818966	21.1361679	0.7025435
dqn_semantic_state	0.1846151	0.7457376	25.3275609	0.4564798	20.7058817	0.6966453
galago_baseline	0.1484205	0.6008163	24.1543098	0.452551	18.9989242	0.6942974
ictnet_div_qe	0.1682738	0.6813918	26.6007845	0.4878239	20.9705288	0.6942803
ictnet_emulti	0.1682738	0.6813918	26.6007845	0.4878239	20.9705288	0.6942803
ictnet_fom_itr1	0.1679748	0.6799596	26.8963205	0.4887893	22.5396835	0.6968833
ictnet_params1_s	0.1691698	0.6854357	26.5121075	0.4867414	21.0638491	0.6945301
ictnet_params2_ns	0.159277	0.6445518	26.8218567	0.4943407	20.9017879	0.6938016

Table 6. Iteration 4

	CT	nCT	sDCG	nsDCG	EU	nEU
clip_addwords	0.0678964	0.3444249	16.8335322	0.2948765	8.726744	0.7092483
clip_baseline	0.1228053	0.6214677	25.9919143	0.4492134	19.672292	0.7298892
clip_filter	0.108693	0.5521342	14.0906933	0.2341365	6.5447232	0.7053321
dqn_5_actions	0.1242476	0.6294138	28.1966415	0.4836972	21.0581117	0.7395657
dqn_semantic_state	0.169166	0.8536985	25.8106292	0.4440257	20.3719695	0.7338354
galago_baseline	0.1211844	0.6132961	25.3420453	0.450617	19.1343022	0.7323573
ictnet_div_qe	0.1535742	0.7775634	27.7020265	0.4836873	22.0125427	0.7340171
ictnet_emulti	0.1535742	0.7775634	27.7020265	0.4836873	22.0125427	0.7340171
ictnet_fom_itr1	0.1515121	0.7672265	28.0996203	0.4821021	23.571531	0.7361476
ictnet_params1_s	0.1534574	0.7776179	27.634994	0.4799675	21.9731633	0.7339534
ictnet_params2_ns	0.129882	0.6569213	28.4645202	0.49818	21.5761693	0.7326871

Table 7. Iteration 5

	CT	nCT	sDCG	nsDCG	EU	nEU
clip_addwords	0.0678964	0.4133099	16.8335322	0.2828022	8.726744	0.7406976
clip_baseline	0.1228053	0.7457613	25.9919143	0.4318069	19.672292	0.7588699
clip_filter	0.108693	0.6625611	14.0906933	0.2251995	6.5447232	0.7372452
dqn_5_actions	0.1114354	0.6770347	29.2426583	0.4840687	21.3944172	0.7682166
dqn_semantic_state	0.1589661	0.9625451	26.1356308	0.4332171	20.1245082	0.7620878
galago_baseline	0.102758	0.6240457	26.5015414	0.451918	19.1694004	0.7608299
ictnet_div_qe	0.1482833	0.9013301	28.2219353	0.4731037	22.5937155	0.7632705
ictnet_emulti	0.1482833	0.9013301	28.2219353	0.4731037	22.5937155	0.7632705
ictnet_fom_itr1	0.1457767	0.8862592	28.6406642	0.471537	24.1143185	0.765135
ictnet_params1_s	0.1479293	0.8997561	28.0461698	0.4682488	22.2423695	0.7626915
ictnet_params2_ns	0.1088391	0.6605537	29.3461089	0.495446	21.0809801	0.7601184

Table 8. Iteration 6

	CT	nCT	sDCG	nsDCG	EU	nEU
clip_addwords	0.0678964	0.4821948	16.8335322	0.2741452	8.726744	0.7652617
clip_baseline	0.1228053	0.8700548	25.9919143	0.4192483	19.672292	0.7815426
clip_filter	0.108693	0.7729879	14.0906933	0.2188205	6.5447232	0.7621639
dqn_5_actions	0.1014448	0.7187933	29.9681515	0.4836611	21.2873622	0.7899146
dqn_semantic_state	0.1532954	1.0830598	26.4518386	0.425639	19.9831539	0.784234
galago_baseline	0.0892321	0.6321475	27.6992079	0.4548045	19.6283697	0.7840226
ictnet_div_qe	0.1471464	1.0435935	28.4051434	0.4624315	22.6991193	0.7856276
ictnet_emulti	0.1471464	1.0435935	28.4051434	0.4624315	22.6991193	0.7856276
ictnet_fom_itr1	0.1436776	1.0192755	28.8950393	0.4613288	24.3585699	0.7875192
ictnet_params1_s	0.145998	1.0361967	28.2125682	0.457354	22.3501344	0.7851373
ictnet_params2_ns	0.0947483	0.6708001	30.1182616	0.4932847	20.2910964	0.7812922

Table 9. Iteration 7

	CT	nCT	sDCG	nsDCG	EU	nEU
clip_addwords	0.0678964	0.5510798	16.8335322	0.2673327	8.726744	0.7851366
clip_baseline	0.1228053	0.9943484	25.9919143	0.4094087	19.672292	0.7999115
clip_filter	0.108693	0.8834147	14.0906933	0.2138184	6.5447232	0.782321
dqn_5_actions	0.094049	0.7613281	30.6252363	0.4857005	21.1754035	0.8075685
dqn_semantic_state	0.1496234	1.208232	26.6684806	0.4182104	19.6949104	0.8018887
galago_baseline	0.0792522	0.6417305	28.4367084	0.4564056	18.9915583	0.8011377
ictnet_div_qe	0.1464791	1.18734	28.5742948	0.45413	22.9151812	0.8038672
ictnet_emulti	0.1464791	1.18734	28.5742948	0.45413	22.9151812	0.8038672
ictnet_fom_itr1	0.1430616	1.1599584	28.9817328	0.4510654	24.4402263	0.8054902
ictnet_params1_s	0.1454838	1.1801108	28.3264534	0.4481567	22.4966348	0.8033499
ictnet_params2_ns	0.0846225	0.6845636	31.0150758	0.4943328	19.8152508	0.7987967

Table 10. Iteration 8

	CT	nCT	sDCG	nsDCG	EU	nEU
clip_addwords	0.0678964	0.6199648	16.8335322	0.2620276	8.726744	0.8017508
clip_baseline	0.1228053	1.1186419	25.9919143	0.4017303	19.672292	0.8152839
clip_filter	0.108693	0.9938416	14.0906933	0.2099336	6.5447232	0.7991677
dqn_5_actions	0.0877909	0.7992842	31.197493	0.4852854	20.9258899	0.8220898
dqn_semantic_state	0.1475708	1.3405783	26.9146451	0.4124845	19.7538581	0.8171259
galago_baseline	0.0710383	0.6471881	29.2236247	0.4592193	18.7172524	0.8159354
ictnet_div_qe	0.1460904	1.3322595	28.6543297	0.4462176	22.9764141	0.8190047
ictnet_emulti	0.1460904	1.3322595	28.6543297	0.4462176	22.9764141	0.8190047
ictnet_fom_itr1	0.1428165	1.302747	29.0845744	0.4432499	24.5779679	0.820589
ictnet_params1_s	0.1452148	1.3252039	28.4533857	0.4411333	22.6637688	0.8186212
ictnet_params2_ns	0.0763365	0.694617	32.058902	0.5003299	20.0493688	0.8143285

Table 11. Iteration 9

	CT	nCT	sDCG	nsDCG	EU	nEU
clip_addwords	0.0678964	0.6888498	16.8335322	0.2576967	8.726744	0.8158557
clip_baseline	0.1228053	1.2429355	25.9919143	0.3955611	19.672292	0.8283471
clip_filter	0.108693	1.1042684	14.0906933	0.2067762	6.5447232	0.8134679
dqn_5_actions	0.082483	0.834239	31.7663899	0.4850046	20.9282911	0.8347376
dqn_semantic_state	0.1464314	1.4784141	27.1194546	0.4107508	19.8255847	0.8302033
galago_baseline	0.0643294	0.6512856	29.6418512	0.4581143	17.9727267	0.8280541
ictnet_div_qe	0.1458809	1.4781937	28.73881	0.4400128	23.0661807	0.8318932
ictnet_emulti	0.1458809	1.4781937	28.73881	0.4400128	23.0661807	0.8318932
ictnet_fom_itr1	0.1426205	1.4455368	29.1595962	0.4366922	24.6680233	0.8333821
ictnet_params1_s	0.1449997	1.4702983	28.5416048	0.4351139	22.7628184	0.8315481
ictnet_params2_ns	0.0689686	0.6973144	32.8403972	0.5033402	19.1573553	0.826299

Table 12. Iteration 10