



# Open Research Online

---

The Open University's repository of research publications and other research outputs

## Overview of the TREC 2011 Session Track

### Conference or Workshop Item

How to cite:

Kanoulas, Evangelos; Hall, Mark; Clough, Paul; Carterette, Ben and Sanderson, Mark (2012). Overview of the TREC 2011 Session Track. In: The Twentieth Text REtrieval Conference (TREC 2011) Proceedings, 15-18 Nov 2011, Gaithersburg, Maryland, USA.

For guidance on citations see [FAQs](#).

© [\[not recorded\]](#)

Version: Version of Record

Link(s) to article on publisher's website:

<https://trec.nist.gov/pubs/trec20/t20.proceedings.html>

---

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data [policy](#) on reuse of materials please consult the policies page.

---

[oro.open.ac.uk](http://oro.open.ac.uk)

# Overview of the TREC 2011 Session Track

Evangelos Kanoulas\* Ben Carterette† Mark Hall‡ Paul Clough§ Mark Sanderson¶

## 1 Introduction

The TREC Session track ran for the second time in 2011. The track has the primary goal of providing test collections and evaluation measures for studying information retrieval over user *sessions* rather than one-time queries. These test collections are meant to be portable, reusable, statistically powerful, and open to anyone that wishes to work on the problem of retrieval over sessions.

The second year has seen a near-complete overhaul of the track in terms of topic design, session data, and experimental evaluation. The changes are:

1. topics were formed from real user sessions with a search engine, and include queries, retrieved results, clicks, and dwell times;
2. retrieval tasks designed to study the effect of using increasing amounts of user data on retrieval effectiveness for the  $m$ th query in a session;
3. subtopic relevance judgments similar to the Web track diversity task.

We believe the resulting test collection better models the interaction between system and user, though there is certainly still room for improvement.

This overview is organized as follows: in Section 2 we describe the tasks participants were to perform. In Section 3 we describe the corpus, topics, and sessions that comprise the test collection. Section 4 gives some information about submitted runs. In Section 5 we describe relevance judging and evaluation measures, and Sections 6 and 7 present evaluation results and analysis. We conclude in Section 8 with some directions for the 2012 Session track.

## 2 Evaluation Tasks

We use the word “session” to mean a sequence of reformulations along with any user interaction with the retrieved results in service of satisfying an information need. The primary goal for participants

---

\*Information School, University of Sheffield, Sheffield, UK

†Department of Computer & Information Sciences, University of Delaware, Newark, DE, USA

‡Information School, University of Sheffield, Sheffield, UK

§Information School, University of Sheffield, Sheffield, UK

¶Department of Computer Science & Information Technology, RMIT University, Melbourne, Australia

of the 2011 track was to provide the best possible results for the  $m$ th query in a session given data from the session leading up to it.

NIST provided a set of 76 sessions of varying length (described in more detail in Section 3). Each session consists of:

- the current query  $q_m$ ;
- the query session prior to the current query:
  1. the set of past queries in the session,  $q_1, q_2, \dots, q_{m-1}$ ;
  2. the ranked list of URLs for each past query;
  3. the set of clicked URLs/snippets and the time spent by the user reading the corresponding to each clicked url webpage.

Participants then ran their retrieval systems over only the current query under each of the following four conditions separately:

- RL1** ignoring the session prior to this query
- RL2** considering only the item (1) above, i.e. the queries prior to the current
- RL3** considering only the items (1) and (2) above, i.e. the queries prior to the current along with the ranked lists of URLs and the corresponding web pages
- RL4** considering all the items (1), (2) and (3) above, i.e the queries prior to the current, the ranked lists of URLs and the corresponding web pages and the clicked URLs and the time spent on the corresponding web pages

Comparing the retrieval effectiveness in (RL1) with the retrieval effectiveness in (RL2)–(RL4) one can evaluate whether a retrieval system can use increasing amounts of information prior to a query to improve effectiveness for that query.

### 3 Test Collection

Our test collection consists of a corpus, a set of topics, and relevance judgments (described in the next section). But unlike most test collections, ours also includes a set of *sessions* of user interactions (including query reformulations). A single topic can have more than one session associated with it, since two different sessions could go about satisfying the same information need in very different ways and with different degrees of success.

#### 3.1 Corpus

The track used the ClueWeb09 collection. The full collection consists of roughly 1 billion web pages, comprising approximately 25TB of uncompressed data (5TB compressed) in multiple languages. The dataset was crawled from the Web during January and February 2009. Participants were encouraged to use the entire collection, however submissions over the smaller “Category B” collection of 50 million documents were accepted. Note that Category B submissions was evaluated as if they were Category A submissions.

## 3.2 Topics

Topics are defined in the usual TREC ad hoc sense, with a title, description, and narrative. To define a set of topics, we started with TREC 2009 Million Query track queries and TREC 2007 Question Answering track questions. To increase the chances of collecting longer, more interesting sessions, we attempted to select topics that we believed would require more reformulations before the need would be satisfied. In particular, we were interested in faceted topics, i.e. topics that have a clear set of subtopics that correspond to partial information needs (similar to those defined in the Web track’s diversity task [4]). For a user to cover all the partial information needs, they would very likely need to pose a sequence of queries to the search engine, each query covering a different subset of those partial needs.

The TREC Question Answering track used a set of topics that had clear sets of subtopics [5]. We used the 2007 iteration of the track so that topics are closer to the time period of the ClueWeb09 crawl. An example below is for the topic “Paul Krugman”:

QA topic: 216  
Title: Paul Krugman  
Questions: For which newspaper does Krugman write? At which university does Krugman teach? From which university did he receive his doctorate? What is Krugman’s academic specialty? What prize originating in Spain has Krugman won? What are titles of books written by Krugman? What persons have Krugman criticized in his op-ed columns?

We also explored the TREC 2009 Million Query track topics, which consisted of 40,000 queries taken from the log of a commercial search engine [2]. Some of these queries were back-fit into topics by an assessor who wrote a description and narrative describing a possible information need they might represent. These were also classified by NIST assessors into six categories by whether they were “precision-oriented” or “recall-oriented”. We chose those in the “recall” category, and we further ensured that the narrative of the topic included a number of questions asked by the assessor that originally judged for the query. An example can be seen below.

MQ topic: 20419  
Title: how to get a divorce  
Questions: Where does a person start in trying to get a divorce? Is a lawyer the best person to see first? How can the Internet be used in obtaining a divorce? Is it possible to obtain a divorce without seeing a lawyer? How long does it usually take to get a divorce?

Thus each topic formed for the track had a number of subtopics, roughly corresponding to the questions asked in the description field of a Question Answering topic or the narrative field of a Million Query topic. An example of such a topic can be seen below.

```
<topic num="14">  
  <session num="15"/>  
  <session num="16"/>  
  <title>hardees</title>
```

```
<desc>What is the present situation of Hardees?</desc>
<subtopic number="1" type="inf">
  When and by whom was Hardees founded?
</subtopic>
<subtopic number="2" type="inf">
  Where are the corporate offices?
</subtopic>
<subtopic number="3" type="inf">
  Where are most of the restaurants located?
</subtopic>
<subtopic number="4" type="inf">
  What are its main products?
</subtopic>
<subtopic number="5" type="inf">
  Is Hardees affiliated with other companies? If so, which?
</subtopic>
</topic>
```

In this example sessions 15 and 16 (described in more detail below) focused on topic 14. The title of the topic is “hardees”, the general description of the topic is “What is the present situation of Hardees?”, and five subtopics corresponding to five partial information needs are provided.

### 3.3 Sessions

A session is a series of actions, including queries and clicks on ranked results, that a user performs in the process of trying to satisfy the information need represented by the topic. Through the process described above we arrived at a large set of candidate topics for the track. These topics were then presented to actual users (minus the explicit subtopics, but with the narrative including the list of questions), who would see five randomly-selected topics and asked to choose one to try to satisfy. Users then were able to use a fully-functional custom search engine for ClueWeb09 in order to satisfy the information need described by the topic.

The custom-built search interface first provided instructions to users on the tasks to be conducted (Fig. 7 in Section A). At the beginning of each session with the system, the user was shown five topics sampled randomly from the collection of topics. The user was then prompted to select one of the topics and use the search interface to satisfy the information need (Fig. 8). The user was free to input any queries they liked (Fig. 9), see titles and snippets for retrieved results (Fig. 10), click on URLs to see pages, and continue in this way until they determined they were finished with the topic.

The search interface used the Yahoo! BOSS (Build your Own Search Service) API to perform the actual queries, and then filtered the ranked results against the ClueWeb09 collection before they were shown to the user. This guarantees that the URLs returned are in the ClueWeb09 collection. It *cannot* guarantee that the document content is the same; in fact, it is likely that many pages have changed since ClueWeb09 was crawled. This is a compromise that we made in order to have a search system that users were likely to find satisfactory. The system requested results from Yahoo!

BOSS until at least 50 documents matching ClueWeb09 URLs had been found, or 50 result pages had been requested from Yahoo! BOSS (whichever came first).

During the course of interacting with the engine, it recorded the user's interactions with the retrieval system, including the queries issued, query reformulations, items clicked, and mouse movements in the results page. Users were also asked to judge the web pages they clicked with respect to relevance to the information need (Fig. 11). All logged information was anonymous and no identifying information about the users was saved. When the users indicated that they had fulfilled the information need, they were presented with a brief exit survey (Fig. 12) that aimed to quantify how well the search system did.

Users were mostly faculty, staff, and students at the University of Sheffield. We sent a university-wide mailing asking for participation; anyone was free to use the system. The overall approach is similar to that described by Zuccon et al. [8] for "crowdsourcing" interactions.

When data collection was complete, we had acquired a set of candidate sessions to go with the candidate topics we defined above. Each session consists of a topic, a set of queries actual users posed to Yahoo! BOSS about the topic, the returned results and the user interactions with the returned results.

We then performed some automatic and manual culling of sessions to try to achieve a set that would be interesting for the track. This involved eliminating sessions in which the user clearly didn't understand the task or the information need, eliminating sessions in which the need was satisfied after only one query, and preferring sessions with more interactions. When the culling was complete, we had a set of 76 sessions for 62 topics to release to participants.

The sessions were provided in an XML file format. An example session containing all RL4 data might look like this:

```
<session num="1" starttime="08:59:47.258675">
  <topic>
    <title>peacecorp</title>
    <desc>Find information about the peace corp</desc>
    <narr>When was it started and by whom? What services does it provide and
      where does it provide these services? What is the criteria for applying?
      What is the salary or stipend? What positions are available?</narr>
  </topic>
  <interaction num="1" starttime="09:00:04.155323">
    <query>peace corp</query>
    <results>
      <result rank="1">
        <url>http://www.peacecorps.gov/</url>
        <clueweb09id>clueweb09-en0011-60-08003</clueweb09id>
        <title>Peace Corps</title>
        <snippet>Fighting hunger, disease, poverty, and lack of opportunity.</snippet>
      </result>
      ...
      <result rank="10">
        <url>http://www.peacecorpsonline.org/</url>
        <clueweb09id>clueweb09-en0133-67-12909</clueweb09id>
        <title>Peace Corps Online</title>
      </result>
    </results>
  </interaction>
</session>
```

```

        <snippet>Online community of returned Peace Corps volunteers.</snippet>
    </result>
</results>
<clicked>
    <click num="1" starttime="09:00:09.943356" endtime="09:01:13.434255">
        <rank>1</rank>
    </click>
    <click num="2" starttime="09:01:18.582078" endtime="09:02:42.552354">
        <rank>2</rank>
    </click>
</clicked>
</interaction>
<interaction num="2" starttime="09:02:55.569644">
    <query>peace corp apply</query>
    <results>
        ...
    </results>
    <clicked>
        <click num="1" starttime="09:03:02.615239" endtime="09:03:33.507677">
            <rank>2</rank>
        </click>
    </clicked>
</interaction>
<currentquery starttime="09:04:03.469341">
    <query>peace corp application</query>
</currentquery>
</session>

```

Each experimental condition drops more data from the XML format. An RL3 session would include everything except the <clicked> blocks. An RL2 session eliminated the <results> blocks along with the <clicked> blocks. An RL1 session had virtually no information, eliminating entire <interaction> blocks.

There is a median of two reformulations prior to the last query (mean = 2.68). 38% of sessions have three or more reformulations prior to the last query. The maximum number of queries in any session in the set is 10. There are a total of 183 recorded clicks across all sessions (2.4 per session on average). However, there are 12 sessions with no recorded clicks, and therefore an average of 2.9 clicks per session that has at least one click.

## 4 Submissions

Sites were permitted to submit up to three runs. Each submitted run includes four separate ranked result lists for all 76 sessions. Files were named “runTag.RLn”, where “runTag” is a unique identifier for the site and the particular submission, and “RLn” is RL1, RL2, RL3, or RL4 depending on the experimental condition.

The track received 34 runs from the 13 groups listed in Table 1.

---

1.	Bauhaus-Universitt Weimar, Germany
2.	Centrum Wiskunde & Informatica (CWI), Netherlands
3.	Dalian University of Technology, China
4.	Beijing University of Posts and Telecommunications, China
5.	Institute of Computing Technology, Chinese Academy of Sciences, China
6.	RMIT University, Australia
7.	Robert Gordon University, UK Institute of Information Science & Technologies (ISTI), Italy <sup>1</sup>
8.	Rutgers University, USA
9.	University of Amsterdam, Netherlands
10.	University of Delaware, USA
11.	University of Essex, UK Center for Research and Technology, Thessaly, Greece <sup>1</sup>
12.	University of Massachusetts Amherst, USA
13.	University of Pittsburgh, USA

---

Table 1: Groups participating in the 2011 Sessions Track.

Section B at the end of the document summarizes the methods used by each of the participating sites. For further details on the techniques used refer to the individual groups reports for the Session Track.

## 5 Session Evaluation

### 5.1 Relevance Judgments

Judging was done by assessors at NIST. As described above, each topic was the subject of one or more sessions. For each one of the 62 topics, a pool was formed from the ranked results for the past queries  $q_1 \dots q_{m-1}$  produced by Yahoo! BOSS along with the top 10 ranked documents from the submitted runs on the current query  $q_m$  for all corresponding sessions.

The NIST assessors then judged each document in the pool with respect to each one of these subtopics along with the general topic description. Thus, for the Hardees example in the previous section, a document maybe relevant to subtopic 2, describing the location of the corporate offices of Hardees, but irrelevant to the remaining subtopics. Some document may be irrelevant to all the subtopics but may provide some other relevant information about the present situation in Hardees, in which case it is relevant to the general topic.

The qrels produced have the following format:

```
<topic-id> <subtopic> <doc-id> <judgment>
```

For instance:

```
14 0 clueweb09-en0085-94-35427 1
14 1 clueweb09-en0085-94-35427 0
```



```
14 2 clueweb09-en0085-94-35427 0
14 3 clueweb09-en0085-94-35427 1
14 4 clueweb09-en0085-94-35427 1
14 5 clueweb09-en0085-94-35427 0
...
```

where subtopic is 0 for the overall topic description and subtopic number otherwise; and judgment is -2 for spam document (by definition, a document judged as spam must be spam for all subtopics), 0 for not relevant, 1 for relevant, 2 for highly relevant, or 3, which means the topic was navigational in nature and the judged page was “key” to satisfying the need.

A total of 19,413 documents were judged, with a mean of 313 documents in a pool. Out of these 19,413 documents, 469 were judged as key documents for at least one subtopic, 2,257 were judged as highly relevant for at least one subtopic, 5,051 as relevant, 17,915 as non-relevant and 817 as spam. Note that these numbers do not sum up to 19,413 since a document that is relevant to one subtopic may be nonrelevant to another, and thus double-counted.

## 5.2 Relevance of a Document

The current query  $q_m$  of each session over which runs were submitted may correspond only to a particular subtopic of a given topic. For instance, the current query of the session number 1, “peace corp application”, clearly corresponds to the subtopic “Find information about jobs with the Peace Corps, such as criteria for applying, salary/stipend, and available positions.”. The current query may also correspond to more than one subtopic. For instance, the current query of session number 11, “hawaii real estate average resale value house OR condo news” corresponds to both subtopic number 1, “What is the present average resale value for single family houses?”, and subtopic number 2, “What is the average resale value of single family condos?”.

Given this, along with the fact that a document may be relevant to a number of subtopics with different relevance grades for each one of them, there are a number of questions that need to be answered before calculating evaluation measures:

1. Is a document that is relevant to some of the user’s partial information needs (subtopics) but none of them are expressed through the current query still considered relevant?
2. What is the relevance grade of a document relevant to different subtopics with different grade?

We took two extreme approaches for this preliminary evaluation and report:

- (a) we computed one set of evaluation scores by considering relevant those documents that are relevant to any subtopic or the general topic; if a document is relevant to more than one subtopic then the maximum grade is considered as the relevance grade of the document.
- (b) we computed a second set of evaluation scores by considering relevant those documents that are relevant to the subtopic(s) that the current query corresponds to; if a query corresponds to more than one subtopic and a document is relevant to more than one of these subtopics the maximum grade is considered as the relevance grade of the document. The mapping between current queries and subtopics was judged by the co-ordinators of the track and released as part of the track’s test collection.

Furthermore, unlike the 2010 track evaluation, we did not apply any special treatment to duplicate documents, i.e. documents in the ranked lists for the current query that have been returned (and clicked by users) previously in the session.

There are clearly other decisions that could be made. In the future, we intend to compute results for which documents relevant to all subtopics will be considered relevant but those relevant to the corresponding to the current query’s subtopics will be given higher weight. We also intend to apply some diversity measures such as those used by the Web track.

### 5.3 Evaluation Measures

Based on the qrels provided by NIST and the decisions described above, we evaluated the submitted runs by eight measures:

- Expected Reciprocal Rank (ERR) [3]
- ERR@10
- ERR normalized by the maximum ERR per query (nERR)
- nERR@10
- nDCG
- nDCG@10
- Average Precision (AP)
- Graded Average Precision (GAP) [7]

We produced two sets of evaluation scores, one using maximum relevance to any subtopic and one using maximum relevance to any subtopic associated with the current query.

## 6 Evaluation Results

As described above, we performed two separate evaluations. The results are provided here, with the first set in Section 6.1 and the second in Section 6.2.

### 6.1 Evaluation Against All Subtopics

First we look at the evaluation against all subtopics. This will give a sense of whether the extra interaction information provided in each subsequent experimental condition is useful to satisfying the overall information need.

Table 2 shows all results (by nDCG@10) for all submitted runs in all four experimental conditions. If RL1 (no information about the session) is the baseline, about half of the submitted runs were able to improve on that using only the information about prior queries (RL2) or using information about prior queries and retrieved results (RL3). A majority of submitted runs improved on the

run	RL1	RL2		RL3		RL4	
PITTSIS	<b>0.3789</b>	<b>0.4281</b>	↑	0.4282	↑	0.4409	↑
udelWpMnz	0.3743	0.2767	↓	0.3675	↓	0.3904	↑
wildcat2	0.3737	0.3648	↓	0.4098	↑	<b>0.4540</b>	↑
umasscontext	0.3737	0.3392	↓	0.3392	↓	0.4138	↑
DUTIR2011A	0.3677	0.3797	↑	0.3677	↓	0.3676	↓
essexAnchor	0.3634	0.4016	↑	<b>0.4307</b>	↑	0.4175	↑
essexNooNeg	0.3634	0.4195	↑	0.2540	↓	0.2422	↓
essexNooPos	0.3634	0.4195	↑	0.3310	↓	0.3518	↓
rguBase	0.3634	0.3763	↑	0.0000	↓	0.0000	↓
rguPisaSS	0.3634	0.3578	↓	0.3735	↑	0.3759	↑
rguPisaSST	0.3634	0.3578	↓	0.3735	↑	0.3773	↑
umassqdist	0.3627	0.3950	↑	0.3950	↑	0.4014	↑
UvAmodeling	0.3363	0.3363	↔	0.3498	↑	0.4124	↑
webis11ind	0.3351	0.2508	↓	0.2494	↓	0.2579	↓
wildcat1	0.3310	0.3415	↑	0.3654	↑	0.3544	↑
RMIT1	0.3232	0.3028	↓	0.1770	↓	0.2591	↓
RMIT2	0.3232	0.2933	↓	0.1770	↓	0.1872	↓
RMIT3	0.3232	0.3028	↓	0.1770	↓	0.1933	↓
UvAlearning	0.3214	0.3214	↔	0.4000	↑	0.2928	↓
udelASFe1new	0.3201	0.3148	↓	0.3180	↓	0.3871	↑
CWIron1	0.3091	0.3223	↑	0.3160	↑	0.3160	↑
CWipostRW	0.3071	0.3224	↑	0.3224	↑	0.3240	↑
CWIron2	0.3052	0.3215	↑	0.3160	↑	0.4091	↑
<b>baseline</b>	0.3007	–	–	–	–	–	–
UvAsemantic	0.2852	0.3055	↑	0.3554	↑	0.3301	↑
udelBe2	0.2769	0.2904	↑	0.2722	↓	0.3904	↑
wildcat3	0.2400	<i>0.4088</i>	↑	<i>0.4152</i>	↑	<i>0.4200</i>	↑
webis11cnw	0.2344	0.1989	↓	0.1852	↓	0.1728	↓
webis11cnb	0.2270	0.1759	↓	0.1784	↓	0.1536	↓
Rgposneg	0.1965	0.2418	↑	0.3259	↑	0.3407	↑
Rspos	0.1965	0.2418	↑	0.3259	↑	0.3289	↑
Rsposneg	0.1965	0.2418	↑	0.3259	↑	0.3214	↑
ICTNET11SER1	0.1510	0.1452	↓	0.0753	↓	0.0753	↓
ICTNET11SER2	0.1510	0.1432	↓	0.1510	↔	0.1510	↔
ICTNET11SER3	0.1510	0.1432	↓	0.1510	↔	0.1510	↔
mean increase	–	0.0383	29% sgnf.	0.0587	59% sgnf.	0.0676	60% sgnf.
mean decrease	–	-0.0280	27% sgnf.	-0.0834	67% sgnf.	-0.0952	67% sgnf.

Table 2: All results by nDCG@10 for the current query in the session for each condition (sorted in decreasing order of RL1 nDCG@10). Boldface indicates the highest nDCG@10 in the condition. Italics indicate the biggest increase from RL1. ↑, ↓ indicate positive or negative differences from RL1. ↑, ↓ indicate statistically significant ( $p < 0.05$  by a paired two-sided t-test) positive or negative differences from RL1. ↔ indicates no difference from RL1. The **baseline** system is our custom search system described above. The last two rows give the mean increase or decrease and the percentage of increases/decreases that are significant.

baseline using the interaction information (RL4). Though we cannot say for sure that RL1 is a baseline for every submission, it seems *prima facie* reasonable to conclude that the interaction information provided by RL4 can be used to improve automatic retrieval results.

As a side note, it is interesting that there are even improvements from RL1 to RL2, considering that this replicates the 2010 track experiment for which we saw only very small improvements [6]. We hypothesize that participants (many of whom are participating in the track for the second time) have learnt how to perform this task, and also that the data better reflects real user behavior.

Figure 1 shows changes in nDCG@10 from the RL1 baseline (left) or with increasing information (right). The three plots going down the left column show changes in nDCG@10 from using no previous data (RL1) to using greater and greater amounts of previous data. The dashed line is a difference in nDCG of zero; points above that line represent systems that saw an improvement from using the additional data while points below it represent systems that were hurt with the additional data. The 95% confidence intervals give a rough idea of whether the results are significant.

On the right-hand side, Figure 1 shows changes in nDCG@10 with increasing amounts of previous data: going from RL1 to RL2, RL2 to RL3, and RL3 to RL4. A few systems see improvement at every step. This suggests that the extra data really is beneficial for effectiveness.<sup>2</sup>

## 6.2 Subtopic-Specific Evaluation

We next looked at the evaluation against the subtopic(s) corresponding to the current query in the session. This should indicate whether the extra information helps as the information need drifts over the course of the session.

Table 2 shows results (by nDCG@10) for all submitted runs in all four experimental conditions. Again, treating RL1 as the baseline, a little over half of the submitted runs were able to improve using information about prior queries (RL2) or information about queries and retrieved results (RL3). Even more systems were able to improve using all recorded interaction information (RL4). This is particularly interesting considering that previous queries may have represented different subtopics of the information need.

Here we see that fewer of the improvements are statistically significant. We believe this makes sense as well, considering that participants did not know there would be a subtopic-based evaluation and could not tune their systems for it. Table 2 thus may suggest some baseline level of improvement that can be gained from looking at previous interactions without any knowledge of how the information need has drifted.

Figure 2 shows changes in final-subtopic nDCG@10 from RL1 to RL2, RL3, and RL4 (left column), and changes as increasing amounts of interaction information are incorporated from RL1 to RL2 to RL3 to RL4 (right column). Like in Table 3, we see smaller improvements, but still some that are significant.

---

<sup>2</sup>We caution against over-interpreting this, though, as we cannot say for certain whether these conditions are comparable for every submitted run.

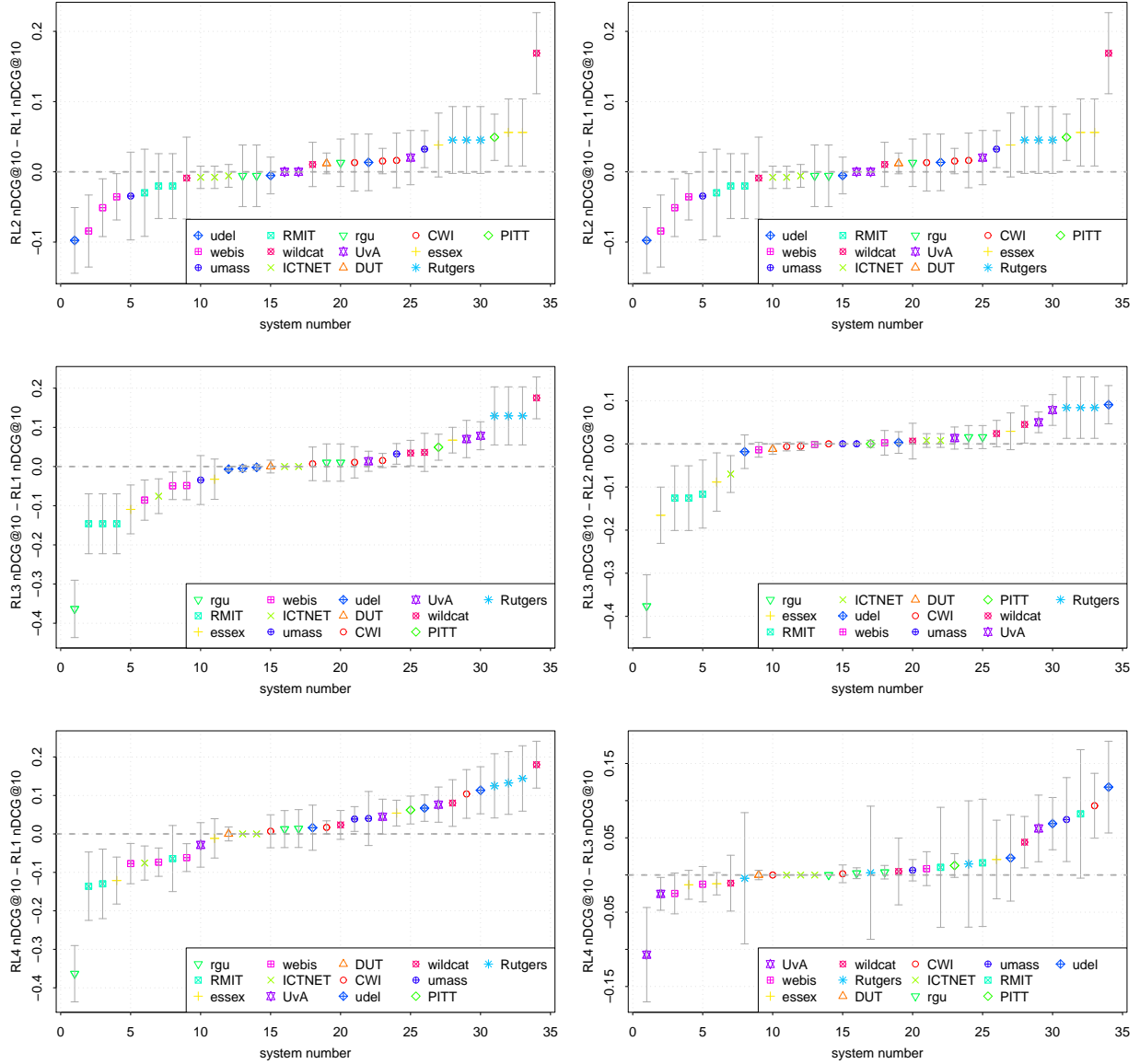


Figure 1: Left: Changes in nDCG@10 from RL1 to (from top to bottom) RL2, RL3, and RL4. Right: Changes in nDCG@10 from RL1 to RL2, RL2 to RL3, and RL3 to RL4. Error bars are 95% confidence intervals.

run	RL1	RL2		RL3		RL4	
udelWPmnz	<b>0.2685</b>	0.1855	↓	0.2644	↓	0.2559	↓
PITTSIS	<b>0.2685</b>	<b>0.2954</b>	↑	<b>0.2981</b>	↑	<b>0.2971</b>	↑
umasscontext	0.2681	0.1888	↓	0.1888	↓	0.2138	↓
umassqdist	0.2605	0.2723	↑	0.2723	↑	0.2736	↑
DUTIR2011A	0.2566	0.2583	↑	0.2495	↓	0.2507	↓
wildcat2	0.2380	0.1975	↓	0.2327	↓	0.2525	↑
UvAmodeling	0.2376	0.2376	↔	0.2328	↓	0.2556	↑
essexAnchor	0.2301	0.2214	↓	0.2710	↑	0.2521	↑
essexNooNeg	0.2301	0.2197	↓	0.1300	↓	0.1219	↓
essexNooPos	0.2301	0.2197	↓	0.1859	↓	0.1913	↓
rguBase	0.2301	0.2259	↓	0.0000	↓	0.0000	↓
rguPisaSS	0.2301	0.2117	↓	0.2064	↓	0.2079	↓
rguPisaSST	0.2301	0.2117	↓	0.2064	↓	0.2109	↓
CWIron1	0.2239	0.1677	↓	0.1627	↓	0.1627	↓
CWIpstRW	0.2219	0.2250	↑	0.2250	↑	0.2285	↑
CWIron2	0.2200	0.1658	↓	0.1627	↓	0.2208	↑
udelASFe1new	0.2187	0.2325	↑	0.2205	↑	0.2448	↑
UvAlearning	0.2112	0.2112	↔	0.2336	↑	0.1732	↓
webis11ind	0.2107	0.1651	↓	0.1656	↓	0.1663	↓
<b>baseline</b>	0.2091	–		–		–	
RMIT1	0.2065	0.1528	↓	0.0984	↓	0.1056	↓
RMIT2	0.2065	0.1352	↓	0.0984	↓	0.0735	↓
RMIT3	0.2065	0.1528	↓	0.0984	↓	0.0751	↓
wildcat1	0.1903	0.1839	↓	0.2084	↑	0.1966	↑
UvAsemantic	0.1863	0.1823	↓	0.2081	↑	0.1927	↑
udelBe2	0.1798	0.2056	↑	0.1759	↓	0.2544	↑
webis11cnw	0.1463	0.1247	↓	0.1173	↓	0.1111	↓
webis11cnb	0.1423	0.1148	↓	0.1063	↓	0.0880	↓
wildcat3	0.1315	<i>0.2324</i>	↑	<i>0.2483</i>	↑	<i>0.2420</i>	↑
Rgposneg	0.1041	0.1224	↑	0.1450	↑	0.1491	↑
Rspos	0.1041	0.1224	↑	0.1450	↑	0.1761	↑
Rsposneg	0.1041	0.1224	↑	0.1450	↑	0.1713	↑
ICTNET11SER1	0.0781	0.0684	↓	0.0338	↓	0.0338	↓
ICTNET11SER2	0.0781	0.0631	↓	0.0781	↔	0.0781	↔
ICTNET11SER3	0.0781	0.0631	↓	0.0781	↔	0.0781	↔
mean increase	–	0.0239	20% sgnf.	0.0324	25% sgnf.	0.0341	27% sgnf.
mean decrease	–	-0.0322	41% sgnf.	-0.0562	41% sgnf.	-0.0667	53% sgnf.

Table 3: Final-subtopic nDCG@10 for the current query in the session (sorted in decreasing order of RL1 nDCG@10). Boldface indicates the highest nDCG@10 in the condition. Italics indicate the biggest increase from RL1. ↑, ↓ indicate positive or negative differences from RL1. ↑, ↓ indicate statistically significant ( $p < 0.05$  by a paired two-sided t-test) positive or negative differences from RL1. ↔ indicates no difference from RL1. The **baseline** system is our custom search system described above. The last two rows give the mean increase or decrease and the percentage of increases/decreases that are significant.

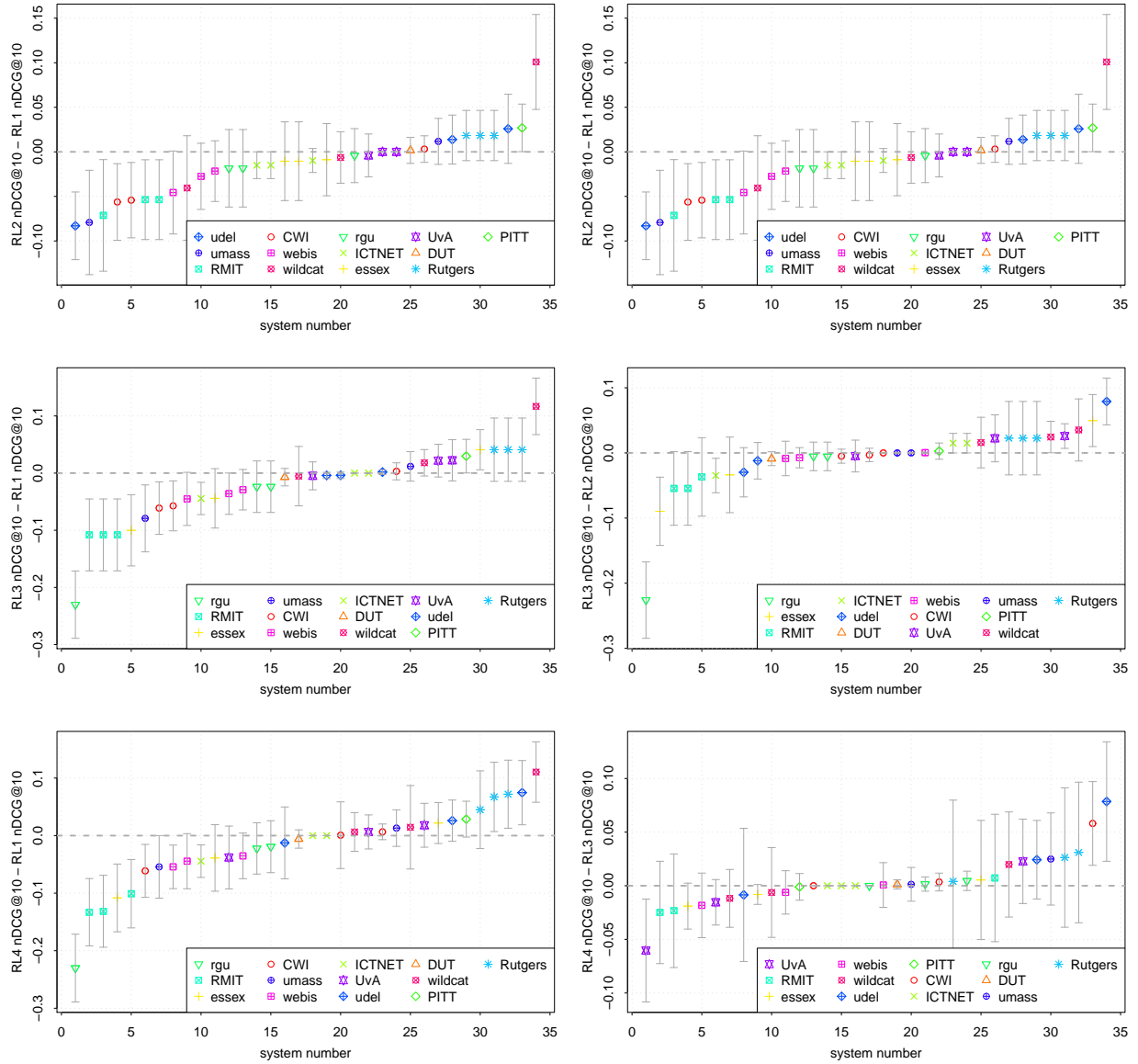


Figure 2: Left: Changes in final-subtopic nDCG@10 from RL1 to (from top to bottom) RL2, RL3, and RL4. Right: Changes in final-subtopic nDCG@10 from RL1 to RL2, RL2 to RL3, and RL3 to RL4. Error bars are 95% confidence intervals.

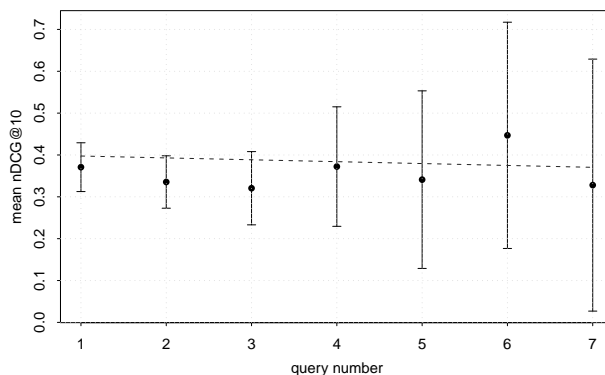


Figure 3: Mean nDCG@10 for our custom search system over the course of user sessions.

## 7 Analysis

For this preliminary report, we picked a few additional questions to investigate to shed some more light on the results.

### 7.1 Evaluating Our Custom System

The RL3 and RL4 conditions allow a system to use information that has come from an altogether different system. In the case of RL3, it is retrieved results, which may encode information about that system’s ranking function. In the case of RL4, it is both retrieved results and user interactions, which may encode information about the utility to users of that system’s ranking function. If our system has a high baseline quality, then, it’s possible that measured improvements in RL3 and RL4 are due to something about our own system rather than something about the evaluated system.

First, we note that it is unlikely that our search system is making any use of interaction information. Since it is API-based, and since we are not sending interaction information back via the API, it could not modify results in light of that information. There is some possibility that it is using the previous query history to rank results, but we have not yet investigated it.

We evaluated our search system by the same official ad hoc evaluation measures used above. Figure 3 shows the mean nDCG@10 averaged over each query in the sessions (e.g. nDCG@10 for query 1 averages the nDCGs for all sessions’ first queries) along with 95% confidence intervals. The confidence intervals increase in width as query number increases because there are fewer and fewer sessions with more queries, from 76 sessions with at least one query down to only two sessions with 8 or more queries (excluded from the figure). With nDCG@10 expected to be between 0.3 and 0.4, it is clear that the system is good, though not substantially better than the top quartile of submitted systems’ RL1 baselines (Table 2). We note, however, that due to differences between our system’s retrieved results and the contents of the ClueWeb collection, these results can only be considered an estimate.

Figure 3 is also suggestive of whether users are picking “better” queries later in the session. It seems that effectiveness remains relatively constant throughout the sessions; while there is a slight downward trend, it is not significant.



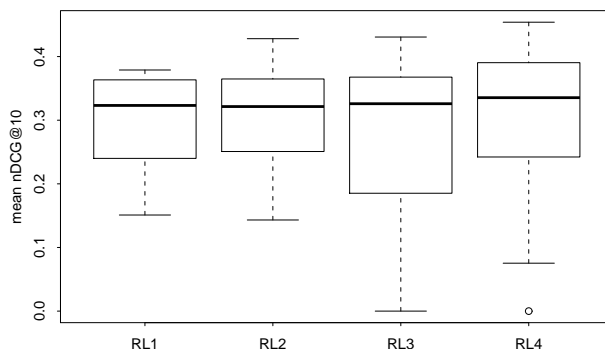


Figure 4: Distribution of submitted runs nDCG@10 for the current query in the session for each RL condition.

Interestingly, for the current query in the session, the effectiveness of our custom system drops significantly to an estimated nDCG@10 of 0.301, which places it in the lower 35% of our submitted runs. This might suggest that users have a tendency to give up once results become sufficiently bad, though it more likely reflects the fact that many of the documents in these rankings were not judged.

Figure 4 shows the distribution of submitted runs nDCG@10 for the current query in the session, for each RL condition separately. The middle bar in each box is median nDCG@10 for the 34 submitted runs. It stays relatively constant over the four RL conditions, whereas if the extra information from the custom system were a significant effect, we would expect to see the median bars rising noticeably. The upper tails of the RL3 and RL4 distributions are certainly higher than those of the RL1 and RL2 distributions, but this would also be expected just from having additional information and cannot be ascribed to specifically having information from a superior system. From this we can tentatively conclude that track participants’ systems are not likely to be taking advantage of information from a better search system.

It could still be the case that only systems that started from a lower baseline than the custom system saw an improvement in their RL3 and RL4 results. There is some evidence to support this: systems that started out from an RL1 baseline close to the custom system saw an average decrease of -0.031 and -0.013 to RL3 and RL4 respectively, while systems that started out from a lower baseline saw an average *increase* of 0.026 and 0.035 to RL3 and RL4 respectively. Nevertheless, there are many individual systems that flout this trend; in particular, CWI, essex, and udel all have at least one system that starts from a high baseline and becomes significantly better.

## 7.2 Correlations Between Evaluation Measures

As mentioned above, in addition to nDCG@10 we also evaluated runs by ERR, AP, and GAP. There are some large differences in results depending on the measure; Table 4 shows rank correlations between nDCG@10, ERR, AP, and GAP for each of the four experimental conditions. Many of these values are quite low for meta-evaluation studies, particularly correlations between nDCG@10 and AP/GAP and between ERR and AP/GAP.

	RL1			RL2			RL3			RL4		
	ERR	AP	GAP	ERR	AP	GAP	ERR	AP	GAP	ERR	AP	GAP
nDCG@10	0.721	0.467	0.523	0.805	0.599	0.603	0.851	0.569	0.609	0.918	0.693	0.686
ERR	–	0.382	0.393	–	0.556	0.538	–	0.543	0.569	–	0.661	0.661
AP	–	–	0.929	–	–	0.960	–	–	0.938	–	–	0.979

Table 4: Kendall’s  $\tau$  rank correlations between four different evaluation measures. Correlations between nDCG@10 and AP measures are quite low compared to “typical” work in meta-evaluation.

One possible reason they are low is simply that there are not many submitted runs to compare. But other TREC tracks have had a low number of submitted runs without having such low correlations between rankings by different measures.

Another possibility is that our systems are not very different from each other, and thus it may not “matter” what order large groups of them are ranked in. The width of the confidence intervals shown in Figure 1 supports this. Many of them overlap to a large degree. This could be because the experiments conducted within the track focus more on improving RL1 baseline results in the RL2, RL3, and RL4 conditions rather than comparing two or more submitted runs to each other directly.

Yet another possibility is that some systems were optimized for different aspects of effectiveness. Figure 5 shows changes in mean average precision over the RL conditions. Comparing to Figure 1 reveals some striking differences: the UvA runs tended to see large and significant improvements in MAP despite not seeing such improvements in nDCG@10, while Rutgers systems that had large and significant improvements in nDCG@10 did not see such improvements in MAP. This lends support to that hypothesis.

### 7.3 Topics With Multiple Sessions

Some of the track’s sessions repeated the same topic. It may be interesting to compare effectiveness between two different sessions trying to satisfy the same information need. For each topic with two or more sessions, for each submitted run, we computed the difference in that run’s effectiveness (by nDCG@10) between the sessions. Figure 6 shows the distribution of differences (over submitted runs) for each of these topics. For some topics the distribution is quite narrow, suggesting similar effectiveness across sessions, but for others it is very wide.

Topic 49 *marsupials*, the subject of sessions 59 and 60, has the widest distribution, with a mean difference in nDCG@10 (for RL4) of 0.037 between the two sessions and a standard deviation of 0.373. These two sessions turn out to be vastly different: the first had only two queries prior to the current one, while the second had 10 queries prior to the current (the most of any session in our set). It is clear that the user in the second session put much more effort into satisfying the information need. Nevertheless, some systems did much better in the first session! For the 18 systems that did better on the second session, the mean difference was 0.336, while for the 16 systems that did better on the first session, the mean difference was 0.384. This suggests that there is a diversity in approaches to using interaction data, and that an approach that works in general may not be best for a particular chain of interactions.

Topic 18 *dooney burke look alike*, the subject of sessions 20 and 21, has a much narrower distribution,

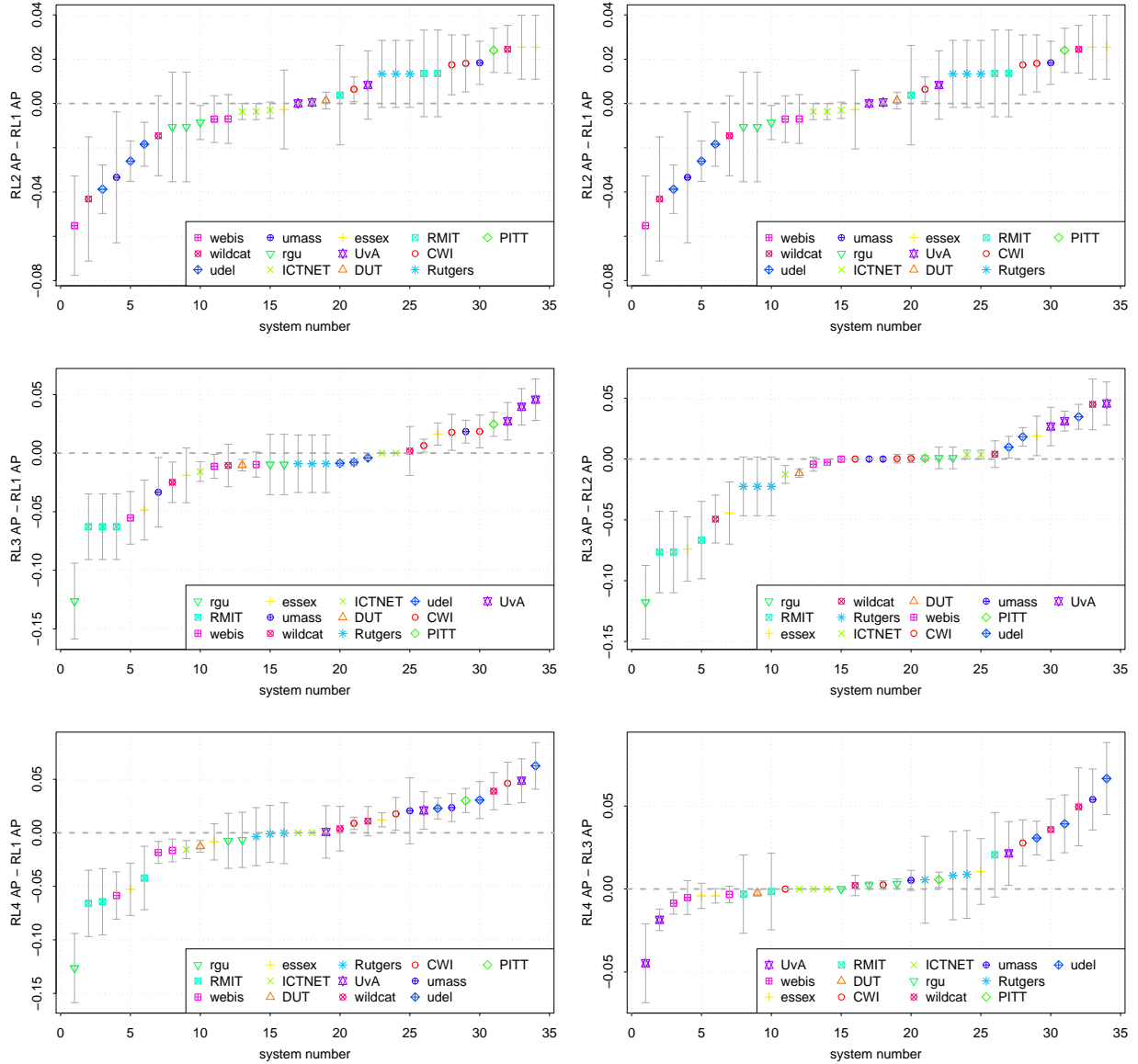


Figure 5: Left: Changes in MAP from RL1 to (from top to bottom) RL2, RL3, and RL4. Right: Changes in MAP from RL1 to RL2, RL2 to RL3, and RL3 to RL4. Error bars are 95% confidence intervals.

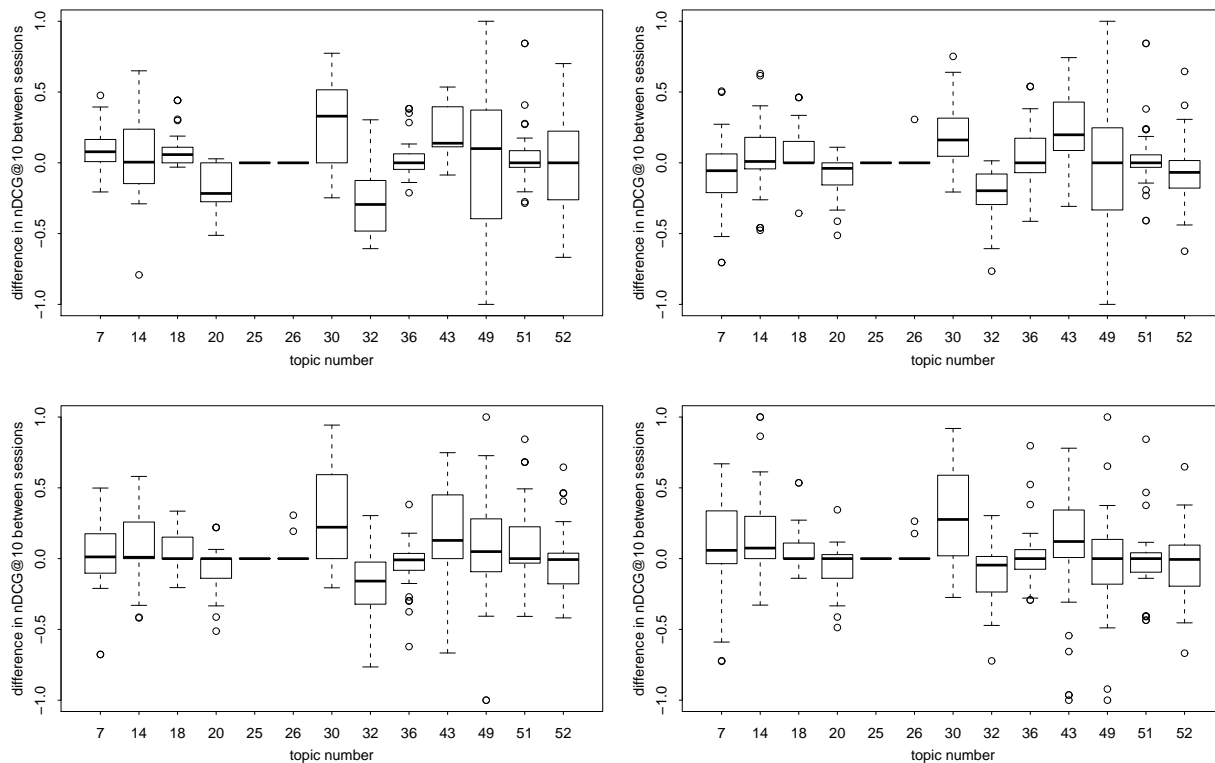


Figure 6: Distributions of differences in nDCG@10 between different sessions based on the same topic. Each distribution is represented as a box plot computed from the submitted runs’ nDCG@10s in (clockwise from upper left) RL1, RL2, RL3, and RL4.

with a mean difference of 0.072 and a standard deviation of 0.145. These two sessions are much more similar to each other, not only in length but also in terms chosen by users for the queries. It is therefore not surprising that effectiveness is similar as well. There is however better effectiveness on session 21 than on session 20; this may be because the user in session 21 used higher-idf terms like “imitation” and “replica” while the user in session 20 used the less informative terms “look alike”.

## 8 Conclusion

The 2011 Session track was a near-complete overhaul of the track, with new tasks, new data, new evaluation methodology, and a new test collection. Preliminary results suggest that it is possible for systems to use provided interaction data to improve automatic results over a baseline that uses no interaction data at all.

There are still questions to explore regarding the evaluation, specifically how to make use of subtopic judgments and how to treat duplicates. These questions will also influence the design of automatic systems for this task.

Looking ahead to the 2012 track, we will likely focus on acquiring more sessions for our topics, so

that we have more than one session per topic on average. This will allow participants to study problems such as different weightings of interactions from different users.

## References

- [1] B. Carterette, E. Kanoulas, P. D. Clough, and M. Sanderson, editors. *Proceedings of the ECIR 2011 Workshop on Information Retrieval Over Query Sessions*, Available at <http://ir.cis.udel.edu/ECIR11Sessions>.
- [2] B. Carterette, V. Pavlu, E. Kanoulas, and F. Hui. Million query track 2009 overview. In E. M. Voorhees and L. P. Buckland, editors, *The Eighteenth Text REtrieval Conference Proceedings (TREC 2008)*. National Institute of Standards and Technology, December 2010. To appear.
- [3] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *In Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, 2009.
- [4] C. L. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2010 Web track. In *Proceedings of the 19th Text Retrieval Conference (TREC)*, 2010. to appear.
- [5] H. T. Dang, J. Lin, and D. Kelly. Overview of the trec 2006 question answering track. In *In Proceedings of the Text REtrieval Conference*, 2006.
- [6] E. Kanoulas, B. Carterette, P. Clough, and M. Sanderson. Session track overview. In *Proceedings of the 19th Text REtrieval Conference (TREC)*, 2010. To appear.
- [7] S. E. Robertson, E. Kanoulas, and E. Yilmaz. Extending average precision to graded relevance judgments. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 603–610, New York, NY, USA, 2010. ACM.
- [8] G. Zuccon, T. Leelanupab, S. Whiting, E. Yilmaz, J. Jose, and L. Azzopardi. Crowdsourcing interactions: Capturing query sessions through crowdsourcing. In Carterette et al. [1].

## A Screenshots of the Search Interface

## B Descriptions of Submitted Runs

Each of the methods used by each one of the groups that participated in the track is summarized below.<sup>3</sup> For further details on the techniques used refer to the individual groups reports for the Session Track.

---

<sup>3</sup>The description of the systems were either taken from the description submitted by participants when submitting the runs or from the participants description sent to the co-ordinates as a proposal for speaking at TREC.

ClueWeb Search Instructions

A set of tasks will appear on the next screen. Please read the tasks carefully and select the task that you are most familiar with and understand best.

After clicking on the "Select this task" button for the task that you wish to select, a search box will appear. Type in the keywords that you believe are the best to find web pages that will help you fulfill the selected task. Please refer to the task description on the left-hand side any time you need to review the task.

Use the search system naturally, as you would do in your everyday searching activities. Browse the ranked list of web-pages, click on any web-page that you think may be useful, click on the next page button if you want to see more results, or reformulate your search query and search again if you think that the returned results are not satisfying.

When viewing a web-page please look at its content, but do not click on any links in it. After viewing a web-page, click the back button to move back to the results page. Before continuing your search activities, please rate the web-page you just viewed. Give a rating of

Perfect	if the web page contains all the information you are looking for and there is no need to consider any other web page,
Excellent	if the web page contains most of the information you are looking for,
Good	if the web page contains just some of the information you are looking for,
Fair	if the web page is relevant to what you are looking for but does not really contain much of the information you are looking for and
Bad	if the web page contains nothing related to what you are looking for.

The rating box will occasionally appear before after clicking on the link to the web-page in the results list, but before you actually see the web-page. If this happens, please ignore the rating box until you have viewed the web-page.

You can finish the task by clicking on the "Finish" button in the left-hand side-panel. You have at most 5 minutes to complete the task but you can stop earlier if you are satisfied with the web pages you have observed so far.

After you finish a task, you will be shown a short questionnaire. Please fill this out and click on the "Save" button. You will then be taken back to the task selection page and can select another task to do. Please do as many tasks as you can in the time that you have.

Note that your browsing activities will be recorded. The recored data is anonymised and then stored securely. It will be used for research purposes only.

Note that the search engine is still experimental and thus slower than most current search engines and will also return fewer results.

There is a help icon (help) in the top-right corner of every page that will show these notes at any time.

If you have any further questions regarding the tasks, please contact e dot kanoulas at sheffield dot ac dot uk. If you find a bug or the system just doesn't work for you, please contact m dot mhall at sheffield dot ac dot uk.

Figure 7: The instruction text used for the data-generation task

**1. Bauhaus-Universitt Weimar, Germany:** The Webis group from the Bauhaus University Weimar submitted three runs, `webis11cnb`, `webis11cnw`, and `webis11ind`. For one run they used the CMU Indri ClueWeb search engine, their two other runs used their own BM25F ClueWeb search engine (all on category A; all with a removal of results with more than 30% spam likelihood).

In their two more mature runs (one with their BM25F engine and one with the Indri search) they used the last query as is for RL1.

For RL2 they derived a maximum (or promising) query that is a query containing as many keywords from the session as possible and still returning about 100 results if possible. Otherwise (if all the keywords together yield more than 100 results) they just use the top 100 results of that query. They

### ClueWeb Search Tasks



Please select one of the five tasks by clicking on the "Select this task" button.

<b>I am looking for horticultural information on black currants.</b> Horticultural information including how to grow black currants is relevant. Health benefits of black currants, products and recipes are not relevant.	Select this task
<b>I am looking for information on the plant Eriogonum jamesii.</b> Information on the culture, growing, uses of Eriogonum jamesii are relevant. Other varieties of Eriogonum are not relevant.	Select this task
<b>Information on parking in San Diego, California</b> Information about parking in San Diego is relevant. References to parking on individual hotel sites are not relevant.	Select this task
<b>Information on Attention Deficit Disorder (ADD).</b> Descriptions and treatment of ADD are relevant. Links to resources for families or individuals dealing with ADD are relevant.	Select this task
<b>I want information or instructions for safety in a roundabout.</b> I want information about or instructions for safety in a roundabout. It can be for any country.	Select this task

Figure 8: The task selection UI.

### Current task

#### Information on Attention Deficit Disorder (ADD).

Descriptions and treatment of ADD are relevant. Links to resources for families or individuals dealing with ADD are relevant.

Finished

### ClueWeb Search

Search

Figure 9: The initial search UI.

**Current task**

**Information on Attention Deficit Disorder (ADD).**

Descriptions and treatment of ADD are relevant. Links to resources for families or individuals dealing with ADD are relevant.

Finished

---

**ClueWeb Search** ?

2008

---

[Dental Website Design, Dental Web Design, Dentist Web Site ...](#)  
 Websites for Dentists. PBHS Leader in Dental Website Design since 1996. Dentist Web design, Dental Practice Marketing, Custom Websites Solutions for Dentists.  
<http://www.pbhs.com/>

[Sturgis.com](#)  
 Sturgis Motorcycle Rally site features lots of photos of the event, as well as schedules, lodging, merchandise, and other information.  
<http://www.sturgis.com/>

[Nobelprize.org](#)  
 Official page from the Nobel Foundation, with a biography of Alfred Nobel, the history and management of the Foundation, information about past and present Laureates, ...  
<http://nobelprize.org/>

[Post Expo 2011](#)  
 Details of the International postal technology exhibition held each year in a different country.  
<http://www.postexpo.com/>

[meetBSD California 2010](#)  
 meetbsd, bsd, unix, freebsd, pcbsd  
<http://www.meetbsd.com/>

Figure 10: The result presentation UI.

Nobel

ear in

ategic

S ...

**Rate page**

Please rate how useful the web-page you just saw was for your current task:

Perfect  
 Excellent  
 Good  
 Fair  
 Bad

Figure 11: The page rating UI.



## ClueWeb Search Exit Questionnaire

Please answer the following questions about the task you just completed:

### How familiar were you with the search topic?

Completely unfamiliar      Very familiar

### How satisfied were you with the quality of the search results?

Completely unsatisfied      Very satisfied

### How satisfied were you with the number of relevant results you found?

Completely unsatisfied      Very satisfied

Save

Figure 12: The exit interview UI

demonstrate in their ECIR 2011 and ICTIR 2011 papers on the User-over-Ranking hypothesis why using maximum queries might be a good idea. The terms of the maximum query were weighted with a discount for terms that only appeared in previous queries and a bonus for terms that only appear in the last query.

For RL3 they used the same query and weighting as with RL2 but excluded all results that appeared in the results presented to the user during previous queries. They also extracted the two most important head noun phrases from the snippets of the presented results from the previous queries and removed all results from the list for the maximum query that contained these head noun phrases.

For RL4 they used the query and weighting of RL2 but also expanded the query with the two most important head nouns of all the clicked snippets of previous queries. From the obtained result list they removed all results presented for previous queries and all the results that contain the two most important head nouns from the not-clicked snippets.

A third run (their baseline) on their BM25F search engine used no term weighting and no head noun extraction of any form from the snippets. But it also used the maximum query idea.

**2. Centrum Wiskunde & Informatica (CWI), Netherlands:** The CWI group submitted three runs, CWIpostR, CWIrun1, and CWIrun2 all on category A. The submissions are based on the observation that a user's ability to issue an effective query in one final step at the end of a session highly depends on how successful she was on pursuing this path. Given that a number of questions arise. Can session log data provide evidences of this procedure's success? And if yes, can they adapt a system's ranking model to its predicted outcome?

A first goal of their experimental trials is to investigate whether session information allows reducing

a possibly very complex process that unfolds when a user interacts with a system, often revising a preliminary information need upon inspection of the results of intermediate queries, to a quality scale between two opposite boundaries. On the one side they find search sessions that are characteristic of ‘good’ users. Good users take on highly selective terms in their queries (RL2), their intermediate result sets show little overlap, meaning that they are effectively investigating different topics’ facets or system’s features (RL3) and they are proficient in inspecting those sets by clicking on the most informative documents (RL4). These users barely benefit from taking too much search history into account to expand their queries: at any point in time the latest queries of good users are nearly optimal.

By contrast there are also ‘bad’ users. Their query terms are not informative enough. Intermediate queries, even when they may seem superficially different, turn out to produce very similar ranked lists. Clicked documents have little impact on their queries afterwards. They forecast that those are the users who benefit most from context-aware systems and even more of suggestions on how to issue better queries.

Assumed that they can single out good and bad users, they also investigate how to deploy a predicted user’s skill into an alternative query representation at the last step of a search session. Experimental results in the behavioural sciences suggest that an exponential discount function represents the effect of past experiences on later behaviour, that the past enters this model as a discount rate rather than as an absolute multiplier and that positive and negative experiences have similar, but opposite effects on a discount rate.

They adapt this formalism to the case of information retrieval. They assume that session records tell success or failure stories of users who became competent questioners, given a topic and a search system, or went astray: a search experience is poised to be rewarding for a ‘good’ user, while the experience of a ‘bad’ user will be negative. They adjust an exponential discount model to the expected quality of a search experience, based on the session information. They use this model to generate a set of weights for terms from past queries, terms from intermediate ranked lists and terms from clicked documents, yielding an alternative representation of the last query in a session.

**3. Dalian University of Technology, China:** The DUTIR group submitted a single run, DUTIR2011A on ClueWeb09 category B.

**4. Beijing University of Posts and Telecommunications, China:** The WILDCAT group submitted three runs, `wildcat1`, `wildcat2`, and `wildcat3` on ClueWeb09 category B. Spam filter was used in all runs. For RL1 the current query was searched using indri language. Some submissions combined pagerank as well. For the RL2 past queries were used to reform the current query and then indri was used on the reformulated query. For RL3 anchor log was used to reform current query, search it in indri, then calculate the similarity between current query and documents. Last for RL4 they use the past queries and the clicked url titles to reform the current query, search it in indri, then calculate the similarity between current query and documents. Alternatives to this included using past clicked urls and their time to calculate similarity with the current search documents and using past clicked urls and time to calculate the similarity between clicked documents and search documents, then predict the time for search documents.

**5. Institute of Computing Technology, Chinese Academy of Sciences, China:** The ICNET group submitted three runs, ICTNET11SER1, ICTNET11SER2, and ICTNET11SER3 on ClueWeb09 category A.

Regarding the first submitted run, for RL1 they apply bm25 on content field, for RL2, they make an intersection from the previous query lists to optimize the result, for RL3 they use the snippet information from xml file to sort the data, then repeat using the intersection, and for RL4 they use the clicked list data to sort the data from the third data. The second run is similar to the first one with different parameters. Regarding the third one, for RL1 they again apply bm25 on content field, for RL2 they use Session Type method and intersection ranking method to optimize the result, for RL3 they use the snippet information from the xml file to classify the primary results from their own search engine and then repeat the session type and intersection rank method to optimize the result further. For RL4, first they use the description and narration information of each query to compute match score with content, then they repeat the procedures based on the third run.

**6. RMIT University, Australia:** The RMIT group submitted three runs, RMIT1, RMIT2, and RMIT3 on category B. The first one served as a baseline: for RL1 just used the current Query, for RL2 using the current Query and all queries in interactions, for RL3 they use query expansion from top five results in all interactions, and for RL4 they use expanded queries from clicked document in all interactions. The next two runs mostly differ in the RL4. The second uses expanded queries from clicked snippets in all interactions and the third is using expanded queries which is used clicked and non clicked results snippets.

**7. Robert Gordon University, UK:** The Robertson Gordon University group in collaboration with the University of Essex and the Institute of Information Science & Technologies (ISTI), Italy, submitted three runs, `rguBase`, `rguPisaSS`, and `rguPisaSST` on ClueWeb09 category B. They propose to use query recommendations extracted from the logs of a web search engine to solve the session track task. The runs are obtained by using the Search Shortcuts recommender system (Broccolo, Marcon, Nardini, Perego & Silvestri 2011). The Search Shortcuts technique uses an inverted index and the concept of satisfactory sessions present in web search engines query log in order to produce effective recommendations for both frequent and rare/unseen queries. They adapt the above technique to work as a query expansion tool. They use such expansion tool to expand the TREC queries; expansion is generated by using a method which aims to consider all past queries in the session. The expansion terms obtained are then used to build a global, uniformly weighted, representation of the user session (RL2). Furthermore, the expansion terms are then combined with ranked lists of results in order to boost terms appearing more frequently in the final results lists (RL3). Finally, they also integrate dwell times and the weighting method obtained taking both result lists and clicks into account for assigning weights to the terms to expand the final query of the session. In addition to that, they submitted a baseline run. It is based on the observation that using the term “wikipedia” to expand the query resulted in a better retrieval performance in last years task, session track 2010 (Lungley, Albakour & Kruschwitz 2011). For RL1 they just submit the current query to the Indri search engine. For RL2, RL3, RL4 they expand the current query with the term wikipedia.

**8. Rutgers University, USA:** The Rutgers University submitted three runs, `Rgposneg`, `Rspos`, and `Rsposneg` on category A. They approached the Session Track task as an issue of personalization, based on both the behaviors exhibited by the searcher during the course of an information seeking episode and a classification of the task that led the person to engage in information-seeking behavior. Their basic approach to personalization is to determine aspects of the task or goal which led the person to engage in information seeking behavior. Other important aspects include determination

of the persons degree of knowledge of the task and search topic, and the persons cognitive abilities. They conducted a series of user studies/experiments in which they controlled the tasks, administered psychometric tests to measure two cognitive abilities, and elicited estimates of the participants knowledge of the tasks and topics of the searches they were asked to perform. They logged a large variety of searcher behaviors, ranging from eye-tracking on search engine result pages (SERPs) and content pages to querying and temporal behaviors of various sorts. The analysis of the data thus collected was aimed at finding associations among the different behaviors (their dependent variables) and the task, knowledge and cognitive abilities information they controlled or elicited (their independent variables), and on the influence of the independent variables and the predictive power of the dependent variables on searcher evaluation of the usefulness of documents with respect to the search task.

This analysis resulted in the generation of several models for prediction of document usefulness. The models are used for either query modification using relevance feedback, or search results re-ranking with terms based on similarity with predicted useful documents. Since their prediction models were generated on the basis of a relatively small number of searches, on quite specific and controlled search task types, and on behaviors on both SERPs and content pages, it is not clear how these models will work with the data available for the TREC Session Track. So, their aim in this Track is to discover whether their models will work on these different task types, and with this different type of data, and if not, why not.

They addressed the Session Track tasks as follows. First, they manually classified the 76 Sessions by task facets based on the Session topic descriptions and narratives. This was used for two of their experimental runs, in which the prediction model was specific to each search task type, combined with search behaviors. For the other runs, they used their general prediction models, which are based on various search behaviors, without reference to task type. The prediction of both useful and not useful documents was then used to modify the last query but one in each search session in a standard relevance feedback mode, one run with positive-only feedback, and two with positive and negative relevance feedback, using the Lemur system in remote mode. Their experimental runs are compared to basic Lemur retrieval with pseudo-relevance feedback.

**9. University of Amsterdam, Netherlands:** The University of Amsterdam submitted three runs, *UvAlearning*, *UvAmodeling*, and *UvAsemantic*. The stream of interactions created by a user engaging with a search system contains a wealth of information. For retrieval purposes, previous interactions can help inform us about a user's current information need. Building on this intuition, the contribution to this TREC year's session track focuses on session modeling and learning to rank using session information. The participants from UvA used three complementary strategies that they designed for improving retrieval for a current query using previous queries and clicked results: probabilistic session modeling, semantic query modeling, and implicit feedback.

Probabilistic session modeling operates on the basis of results retrieved with previous user actions. They model a session as a combination of past queries, sets of result clicks, and the current query. For each of the cues they retrieve a list of results, and combine the result lists according to the session model. They derive the combination parameters using a home-grown collection of training sessions. In this way they create a final list of ranked results per current query.

For semantic query modeling they designed and implemented an algorithm that links query terms to Wikipedia articles and obtains a weighted set of synonyms. They leverage redirects and incoming anchor texts of a Wikipedia entry and calculate the probability that terms refer to this particular

Wikipedia article. This yields a probabilistic model of the language use associated with a Wikipedia entry.

Their implicit feedback strategy uses previous result clicks to learn to combine different ranking features. They implement a regression-based learning to rank algorithm that optimizes linear weighted combinations of ranking features. They train the algorithm on pair-wise labels extracted from result clicks on all but the target session, and on previous queries and results on the target session. Features include a small number of standard retrieval runs and runs implementing the probabilistic session modeling and semantic query modeling discussed above.

**10. University of Delaware, USA:** The University of Delaware submitted three runs, `udelASFe1new`, `udelBe2`, and `udelWPmnz` on category A. They investigated query expansion using clicked and/or unclicked documents retrieved for previous queries: what is the effect on effectiveness when expanding the last query using terms appearing in clicked documents from previous queries; what is the effect when expanding the last query using unclicked documents; and, as a baseline, what is the effect when expanding using all documents retrieved for previous queries?

They also used the Session track to investigate two other questions:

1. how does a corpus filter (e.g. a spam filter, or a "category B" filter, or a Wikipedia filter) affect effectiveness?
2. is there a way to prune retrieved results to remove documents that have already been seen for previous queries (whether relevant or not) without hurting effectiveness?

**11. University of Essex, UK:** The University of Essex, in collaboration with the Center for Research and Technology, Thessaly, Greece submitted three runs, `essexAnchor`, `essexNooNeg`, and `essexNooPos` on category B. The first run is based on an extension of their anchor log technique they proposed in the previous year. They exploit the anchor logs to derive query expansions that are relevant to user information needs throughout the session. This is done by taking up to ten phrases from the anchor text of the displayed or clicked documents of previous queries in the session.

The second and third run are based on Nootropia, a biologically inspired adaptive modelling technique that has been successfully applied to Information Information Filtering (IF) problems such as news recommendation. With Nootropia, information needs are represented as a weighted and ordered network that may represent an individuals multiple topics of interest. This network profile presentation can evaluate the relevance of an information item to the user interests based on a directed spreading activation model. Nodes (features) in the network that appear in the item are activated and subsequently disseminate part of their initial energy to nodes with larger weight. At the end of this feedforward dissemination process, a single relevance score can be calculated as the weighted sum of the final energies of activated terms. They used Nootropia to represent user information needs throughout the search session. The resulting Nootropia profile is then used to re-rank the documents retrieved for the last query in the session. Two different strategies were followed to re-rank the documents in two separate runs. Their first strategy strives to build a profile of user interests and subsequently rank matching documents higher, conversely, the second tries to model what the user is not interested in and rank matching documents lower. Noun phrases extracted from the snippets of the documents were used as representative features to build the Nootropia network.

**12. University of Massachusetts, Amherst, USA:** The UMASS Amherst CIIR group submitted two runs, `umasscontext` and `umassqdist`. The crawl used was ClueWeb09 category A. The first submitted run uses sequential dependency model over the query for RL1, it expands the

current query with most likely terms from previously submitted queries for RL2, it does the same for RL3, and it expands the current query with most likely terms from previously submitted queries and summaries of clicked documents for RL4. The second submitted run uses Score distribution over queries seeded by current query for RL1, uses score distribution over queries seeded by current query, expanded with most likely terms from previously submitted queries for RL2, does the same for RL3, and uses score distribution over queries seeded by current query, expanded with most likely terms from previously submitted queries and clicked document summaries for RL4.

**13. University of Pittsburg, USA:** The University of Pittsburg submitted a single run PITTSIS on category B. The RL1 result only uses the current query. They apply query likelihood language model and tuned sequential dependence model to current query. Documents with lower than 70 Waterloo spam filter score are filtered out. The RL2 result uses the current query and previous queries. They estimate query term necessity based on previous queries in order to weight query terms and phrases (ordered and unordered). Documents with lower than 70 Waterloo spam filter score are again filtered out. The RL3 result further weights query terms and phrases using pseudo-relevance feedback documents of the current query, mixed with RL2's term necessity weight. Instead of using pseudo relevance feedback documents for weighting terms and phrases, RL4 uses clicked results of all previous queries for weighting, mixed with RL2's term necessity weight.