Overview of the TREC 2013 Session Track

Ben Carterette* Evangelos Kanoulas† Mark Hall‡ Ashraf Bah* Paul Clough§

1 Introduction

The TREC Session track ran for the fourth time in 2013. 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 experimental design of the track was similar to that of the previous two years [4, 5]:

- sessions were real user sessions with a search engine that include queries, retrieved results, clicks, and dwell times;
- retrieval tasks were designed to study the effect of using session data in retrieval for only the mth query in a session.

Changes from last year's track include:

1. a new set of topics;
2. a new corpus (ClueWeb12);
3. a new search system for collecting session data;
4. a small change in the retrieval tasks.

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 present evaluation results and analysis.

*Department of Computer & Information Sciences, University of Delaware, Newark, DE, USA
†Google, Zurich, Switzerland
‡Department of Computing, Edge Hill University, Ormskirk, UK
§Information School, University of Sheffield, Sheffield, UK

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 of the 2013 track was to provide the best possible results for the mth query in a session given data from the session leading up to it.

NIST provided participants with a set of 133 sessions of varying length (described in more detail in Section 3). 87 of these were targeted for evaluation; the remaining 46 could be used for training. Each of the 87 evaluated sessions 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, ..., q_{m-1};
    2. the ranked list of URLs for each past query;
    3. the set of clicked URLs/snippets.

In addition, each user action is accompanied by a time relative to the start of the session.

Participants were to run their retrieval systems over only the current query under each of the following three conditions separately:

RL1 ignoring the session prior to this query

RL2 considering all the items (1), (2) and (3) above for the current session, i.e the queries prior to the current, the ranked lists of URLs and the corresponding web pages, the clicked URLs and the time spent on any interaction

RL3 considering all data in the entire session log (in particular, other user sessions on the same topic)

Comparing the retrieval effectiveness in (RL1) with the retrieval effectiveness in (RL2)–(RL3), one can evaluate the effectiveness of different algorithms for incorporating session information into retrieval.

3 Test Collection

Like most IR test collections, ours consists of a corpus, a set of topics, and relevance judgments (described in the next section). 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 ClueWeb12 collection. The full collection consists of roughly 730 million English-language web pages, comprising approximately 5TB of compressed data. The dataset was crawled from the Web during February and March 2012.

Participants were encouraged to use the entire collection, however submissions over the smaller "Category B" collection of about 35 million documents were accepted. Note that Category B submissions was evaluated as if they were Category A submissions. Two of six participating groups used the Category B collection.

3.2 Topics

To define a set of topics, we followed the 2012 track in attempting to control two facets of search tasks as defined by Li and Belkin [6]: "product" and "goal quality". The "product" facet can reflect *intellectual* or *factual* tasks; intellectual tasks produce new ideas or findings (e.g. learn about a topic or make decisions based on information collected), while factual tasks only involve locating facts, data, and other informational items. An example of such variation (from the 2012 Session track topics) can be viewed below.

Query: dehumidifiers

- Session topic: 35
  "Product" facet: Factual
  Description: You would like to buy a dehumidifier. What are some of the technical specifications you should be looking at? What is the price range for dehumidifiers? What makes one dehumidifier more expensive than another?
- Session topic: 37
  "Product" facet: Intellectual
  Description: You would like to buy a dehumidifier. On what basis should you compare different dehumidifiers?

The "goal quality" facet reflects *specific goal(s)* and *amorphous goal(s)*. This is very similar to the dimension that [3] proposed as well-defined and ill-defined information need. Tasks with specific goals have a well-defined information need, while in tasks with amorphous goals, the information need is ill-defined. Tasks with amorphous goals might require users to redefine the topic or identify specific aspects of the subject themselves. An example (again from the 2012 track) can be viewed below:

Query: Swahili dishes

- Session topic: 38
  "Goal" facet: specific goals
  Description: What are some traditional Swahili dishes? What ingredients do they use to cook them? Are swahili people using any particular herb in their dishes? Could you find these ingredients in your country? Are there any recipes you can find online?

- Session topic: 40
  "Goal" facet: amorphous goals
  Description: One of your friends from Kenya invited you to attend a party in his house and have a taste of traditional swahili dishes. You would like to search and find some information about Swahili dishes.

Combining "product" and "goal quality" facets generates four classes of topics, characterized as follows: a factual task with specific goals is *known-item search*, a factual task with amorphous goals is *known-subject search*, an intellectual task with specific goals is *interpretive search*, and an intellectual task with amorphous goals is *exploratory search*. A complete example of all four combinations can be viewed below:

Query: depression symptoms

- Task: Known-item search
  Description: What is depression? What are the major symptoms of depression? What medications, therapies and other treatments can be used to treat depression symptoms? Who performs therapy and what are the costs? Does health insurance pay for any of the treatments?

- Task: Known-subject search
  Description: You think that one of your friends may have depression, and you want to search information about the depression symptoms and possible treatments.

- Task: Interpretive search
  Description: Depression is a loaded word in our culture. What are the symptoms that could differentiate depression from having just a bad month of excessive emotions? When should one seek help and what kind?

- Task: Exploratory search
  Description: A friend has been complaining for months that she is unhappy with her life. She has also mentioned that she can't easily sleep at nights. You think that she may be suffering from depression. You want to understand if this is the case and how you could assist her in getting some help from medical professionals.

Most of the "known-item search" topics were taken from the 2012 Web track and modified and expanded to cover more questions/aspects than the the original Web track queries did. Topics for the other three types of searches were partly inspired by topics that were still in the news while

topics were being developed (e.g. gun violence, gun control) and partly inspired by topics that, at the time, were of interest to or being explored by the person who was generating the session topics (e.g. road-trip, internet phone service, naturalism vs existence of God). Some questions were formulated differently in order to fit under different task types (e.g. internet phone service appeared in both known-item and interpretive search tasks).

Constructing topics of different task types allows to study both how user interactions different across varying task types and whether/how systems can improve search quality under different session characteristics. We developed a total of 69 different topics across these four categories, though we used only 61 of them.

3.3 Sessions

As describe above, 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. The topics were presented to actual users, who would see three randomly-selected topics and asked to choose one to try to satisfy. Topics were selected for displaying to users randomly according to a distribution inverse to how many times the topic had been selected previously, ensuring that topics selected more frequently would be shown less. After selecting a topic, users were able to use a fully-functional custom search engine for ClueWeb12 in order to satisfy the information need described by the topic.

The search system used an indri index of ClueWeb12 as a backend. Each of the 20 ClueWeb12 segments (ClueWeb12_00 through ClueWeb12_19) was indexed using the Krovetz stemmer and no stopword list. The indexes searched contained only text from title fields, anchor text from incoming links ("inline" text), and page URLs. We chose to index these fields only in an effort to get faster response times.

Each query was plugged into an indri query language template as exemplified below for the query "quitting smoking":

```
#filreq(#less(spam -130)
        #weight(1   #combine(quitting smoking)
                1   #weight(1 #combine(quitting.title smoking.title)
                             1 #uw(quitting smoking).title)
                50  #weight(1 #combine(quitting.inlink smoking.inlink)
                             1 #uw(quitting smoking).inlink)
                0.5 #weight(1 #combine(quitting.url smoking.url)
                             1 #uw(quitting smoking).url)))
```

The first line indicates that we only want documents with a spam score of -130 or less, effectively filtering out 65% of the spammiest documents in the collection at query time (based on Waterloo spam scores). The retrieval score is then computed from four component models: a basic query-likelihood model for the full document representation and three weighted combinations of basic query-likelihood field models with unordered-window within-field models. The "inlink" model was weighted 50 times higher than the title model, and 100 times higher than the URL model. This query template is the product of manual search and investigation of retrieved results.

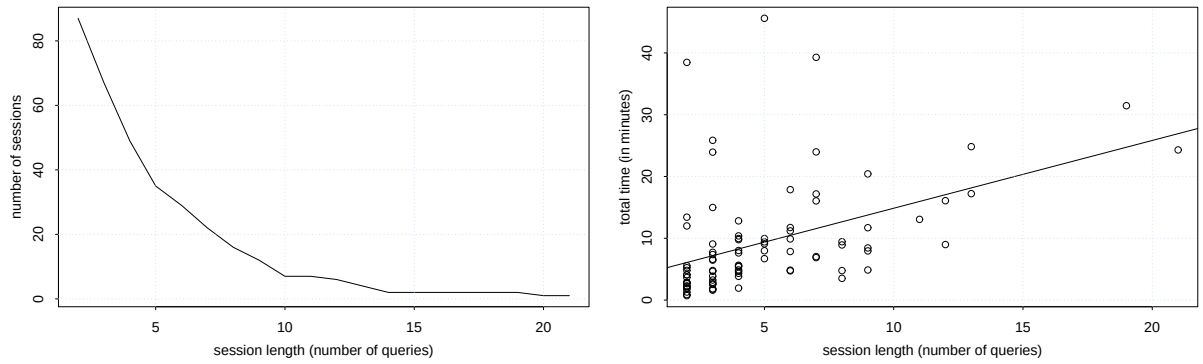

Figure 1: Left: number of sessions of a given length (in terms of total number of queries recorded). Right: amount of time spent in each session.

The search interface connected to our indri backend to retrieve the top 50 results (which were filtered further so that at most two documents from any domain would be shown). Users were shown these results in pages of 10 along with a snippet produced by indri's built-in snippet generator. They could click results to see the current version of the page, which of course could be different from the version in the index.

The system recorded the user's interactions with the retrieval system, including the queries issued, query reformulations, and items clicked in the results page. 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 about the topic, the retrieved results, and the user interactions with the retrieved results.

Session data is provided in an XML file. Part of an example session is shown on page 7.

The released data comprised 133 full sessions. 87 of these have a minimum of one reformulation (two queries total, of which the second is the `currentquery`); these were the ones targeted for evaluation. Of those, 67 have at least three reformulations, 49 have at least four, 29 have at least five, 7 have at least 10, and there is one session with 20 reformulations. Figure 1 shows the distribution of session lengths. The median and mean of the distribution are both 11.5 queries in a session, which reflects much longer sessions than previous years (the mean for 2012 was 2.03). Figure 1 also shows the amount of time spent in the session; the median length was 6.87 minutes. There are also a total of 586 recorded clicks across all 133 sessions, or 4.4 clicks per session on average. This too is higher than previous years.

4 Submissions

Participating sites were permitted to submit up to three runs. Each submitted run includes three separate ranked result lists for all 87 sessions. Files were named "runTag.RLn", where "runTag" is a unique identifier for the site and the particular submission, and "RLn" is RL1, RL2, or RL3, depending on the experimental condition.

```
<session num="10" starttime="0">
   <topic num="12">
      <desc>Your friend would like to quit smoking. You would like to provIde him with
            relevant information about: the different ways to quit smoking, programs
            available to help quit smoking, benefits of quitting smoking, second effects
            of quitting smoking, using hypnosis to quit smoking, using the cold turkey
            method to quit smoking</desc>
   </topic>
   <interaction num="1" starttime="8.30123">
      <query>quit smoking</query>
      <results>
         <result rank="1">
            <url>http://quitsmoking.about.com</url>
            <clueweb12id>clueweb12-0005wb-77-27713</clueweb12id>
            <title>Quit Smoking | Quit Smoking Support | Smoking Cessation</title>
            <snippet>Quit Smoking | Quit Smoking Support | Smoking Cessation
            About.comHealthSmoking Cessation Smoking Cessation Search Smoking
            CessationHealth RisksHow to QuitAfter Quitting Share Quit Smoking
            ToolboxShocking Tobacco Facts10 Things to Avoid When You Quit Guide since
            2003Terry MartinSmoking Cessation GuideSign up for My NewsletterMy Bio...</snippet>
         </result>
         ...
         <result rank="10">
            <url>http://www.heart.org/HEARTORG/GettingHealthy/QuitSmoking/Quit-Smoking_UCM_001085_Su
            <clueweb12id>clueweb12-0300tw-20-20611</clueweb12id>
            <title>Quit Smoking</title>
            <snippet>Quit Smoking ...0 Grams Trans Fat Oils and Fats Restaurant FAQs
            Other Restaurant Resources Quit Smoking Smoking Smoking Life quit today. We
            can help. Learn more. Quit Smoking?Smoking is the most important preventable c
            ause of premature death in the United...</snippet>
         </result>
      </results>
      <clicked>
         <click num="1" starttime="12.984659" endtime="20.557844">
            <rank>3</rank>
         </click>
         <click num="2" starttime="27.030967" endtime="55.220869">
            <rank>1</rank>
         </click>
         <click num="3" starttime="55.220869" endtime="60.704926">
            <rank>6</rank>
         </click>
         <click num="4" starttime="60.704926" endtime="69.165489">
            <rank>5</rank>
         </click>
      </clicked>
   </interaction>
   <currentquery starttime="78.226578">
      <query>quit smoking cold turkey</query>
   </currentquery>
</session>
```

The track received 20 runs from six groups, for a total of 60 ranked lists for each session. They are listed in Table 1.

1.	Bauhaus-Universitat Weimar, Germany
2.	Pattern Recognition and Intelligent System
3.	Georgetown University, USA
4.	Institute of Computing Technology, Chinese Academy of Sciences, China
5.	University of Delaware, USA
6.	University of Pittsburgh, USA

Table 1: Groups participating in the 2013 Sessions 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. Thus pools for judging were formed by topic, not by session.

For each of the 61 topics, a pool was formed from the ranked results produced by our indri system for the past queries $q_1 ... q_{m-1}$ and the current query q_m, along with the top 10 ranked documents from the submitted runs on the current query q_m.The NIST assessors then judged each document in the pool with respect to the topic description. URLs were sorted by domain prior to judging so that assessors would see all pages from the same domain before moving to another one.

The qrels produced have the following format:

```
<topic-id> 0 <doc-id> <judgment>
```

Judgment values are: -2 for spam document (i.e. the page does not appear to be useful for any reasonable purpose; it may be spam or junk.); 0 for not relevant (i.e. the content of this page does not provide useful information on the topic, but may provide useful information on other topics, including other interpretations of the same query); 1 for relevant (i.e. the content of this page provides some information on the topic, which may be minimal; the relevant information must be on that page, not just promising-looking anchor text pointing to a possibly useful page); 2 for highly relevant (i.e. the content of this page provides substantial information on the topic); 3 for key, (i.e. the page or site is dedicated to the topic; authoritative and comprehensive, worthy of being a top result in a web search engine; typically, key pages are more comprehensive, have higher quality, and are from more trustworthy sources than the merely highly relevant page); and 4 for navigational (i.e. this page represents a home page of an entity directly named by the query; the user may be searching for this specific page or site; there is often at most one page that deserves a Navigational judgment for an aspect).

Relevance judgments were eventually transformed to relevance grades with spam and non-relevant documents assigned a grade of 0, relevant assigned a grade of 1, highly relevant assigned a grade of 2, key assigned a grade of 3, and navigational assigned a grade of 4.

A total of 13,132 pages were judged. Out of these 13,132 pages, 44 were judged as navigational, 72 as key, 665 as highly relevant, 2,641 as relevant, 9,591 as nonrelevant, and 113 as spam. On average there were 159 nonrelevant (or spam) and 56 relevant (across all types of relevance) documents per query. Only 14 of the topics had at least one "key" document, and only six had at least one "nav" document.[1]

5.2 Evaluation Measures

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

- Expected Reciprocal Rank (ERR) [2]
- ERR@10
- ERR normalized by the maximum ERR per query (nERR)
- nERR@10
- nDCG
- nDCG@10
- Average Precision (AP)
- Precision@10

6 Evaluation Results

Table 2 shows all results (by nDCG@10) for all submitted runs in all three experimental conditions. If RL1 (no information about the session) is the baseline, nearly all (17 of 20) of the submitted runs were able to improve on that using information about the session prior to the last query (RL2 results). Most of those were statistically significant improvements. A smaller set of systems (7 of 20) were further able to improve by making use of the full log (RL3 results), though only three of these were statistically significant.

Figure 2 shows changes in nDCG@10 from the RL1 baseline (left) or with increasing information (right). The 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 2 shows changes in nDCG@10 with increasing amounts of previous data: going from RL1 to RL2, then RL2 to RL3. Only a few systems see improvement at every step, and the improvements are not significant. There are many possible reasons for this, one of which is simply that the log was simply too small to provide much use beyond the single session.

[1] Somewhat strangely, two topics had more than 15 "key" documents, and one had more than 35 "nav" documents.

run	RL1	RL2		RL3	
wdtiger2	**0.1335**	0.1764	⇑	0.1826	↑
indri baseline	0.1292	–		–	
UDVirtualLM	0.1188	0.1616	⇑	0.1759	⇑
ICTNET13SER2	0.1179	0.1670	⇑	0.1649	↓
ICTNET13SER3	0.1179	0.1617	⇑	0.1608	↓
FixInt28	0.1171	0.1706	⇑	0.1706	↔
FixInt28N	0.1171	0.1663	⇑	0.1663	↔
FixInt58	0.1171	0.1540	⇑	0.1540	↔
FixInt58N	0.1171	0.1521	⇑	0.1521	↔
KM1	0.1171	0.0823	⇓	0.0823	↔
KM1N	0.1171	0.0772	⇓	0.1171	⇑
ICTNET13SER1	0.1170	0.1669	⇑	0.1659	↓
webisS2	0.1114	0.1359	↑	0.1244	↓
webisS1	0.1114	0.1323	↑	0.1130	↓
webisS3	0.1114	0.1256	↑	0.0878	⇓
udelCombUD	0.1058	0.1664	⇑	**0.1940**	⇑
UDVirtualCmb	0.1058	0.0854	↓	0.1009	↑
GUrun3	0.1050	0.1430	⇑	0.1459	↑
GUrun1	0.1043	0.1459	⇑	0.1460	↑
GUrun2	0.1013	0.1480	⇑	0.1485	↑
wdtiger1	0.0941	**0.1952**	⇑	0.1826	↓

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. $\uparrow, \downarrow$ indicate positive or negative differences from the prior condition. $\Uparrow, \Downarrow$ indicate statistically significant ($p < 0.05$ by a paired two-sided t-test) positive or negative differences from the prior condition. $\leftrightarrow$ indicates no difference from the prior condition. The indri baseline system is our custom search system described above.

Figure 3 shows the same information for unnormalized ERR. The two are well-correlated, with linear correlations of 0.85 and higher for all three sets of differences.

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] 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.

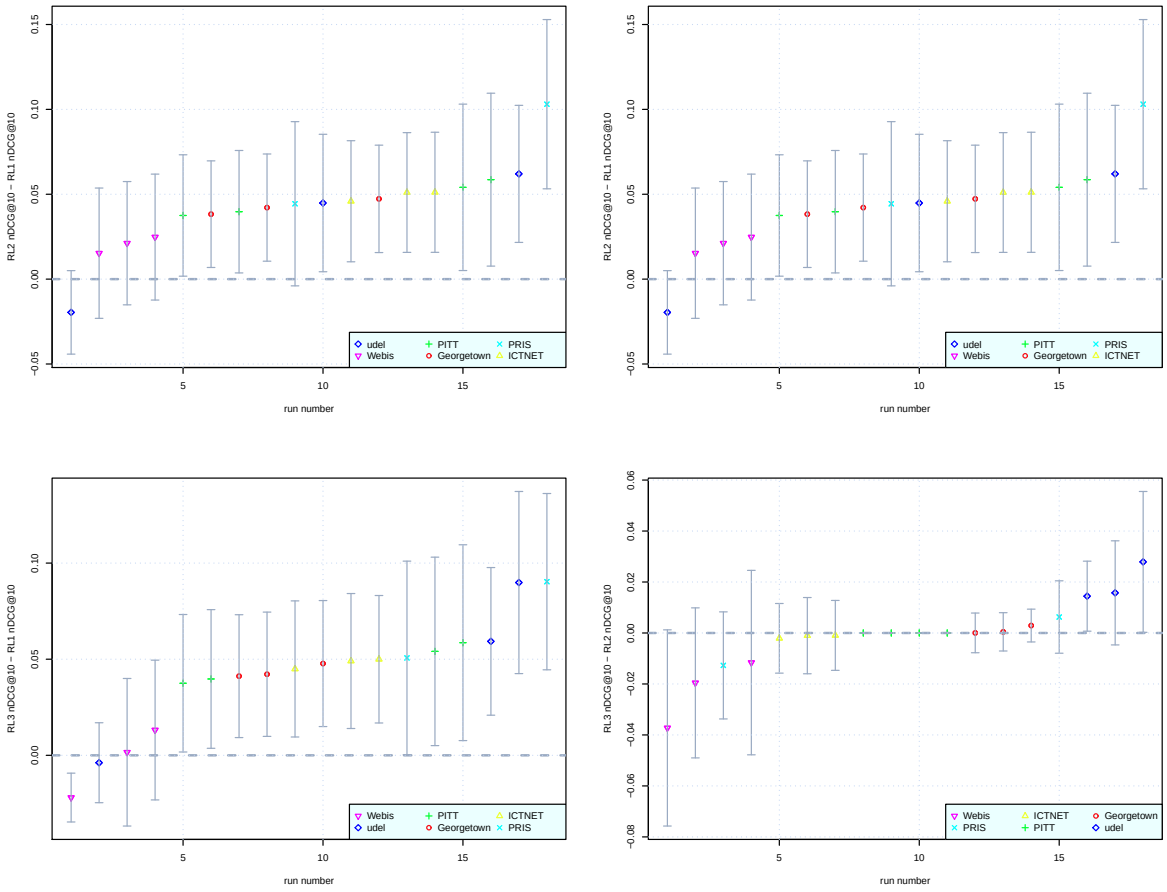

Figure 2: Left: Changes in nDCG@10 from RL1 to (from top to bottom) RL2 and RL3. Right: Changes in nDCG@10 from RL1 to RL2 and RL2 to RL3. Error bars are 95% confidence intervals.

[3] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[4] E. Kanoulas, B. Carterette, M. Hall, P. Clough, and M. Sanderson. Session track 2011 overview. In *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*. National Institute of Standards and Technology, 2012. (`http://trec.nist.gov/pubs/trec20/papers/SESSION.OVERVIEW.2011.pdf`).

[5] E. Kanoulas, B. Carterette, M. Hall, P. D. Clough, and M. Sanderson. Overview of the trec 2012 session track. In *Proceedings of TREC*, 2012.

[6] Y. Li and N. J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *Inf. Process. Manage.*, 44(6):1822–1837, Nov. 2008.

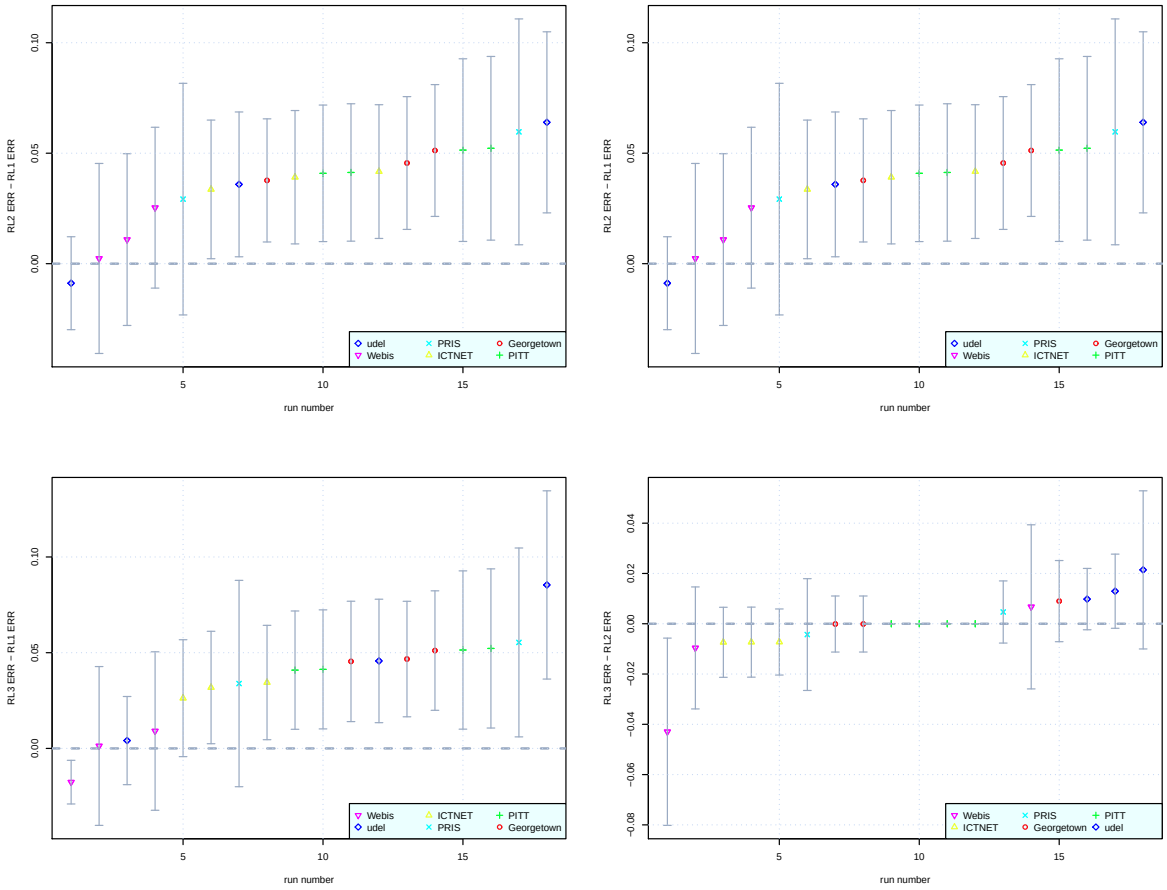

Figure 3: Left: Changes in ERR from RL1 to (from top to bottom) RL2 and RL3. Right: Changes in ERR from RL1 to RL2 and RL2 to RL3. Error bars are 95% confidence intervals.