Evaluating Retrieval over Sessions: The TREC Session Track 2011–2014

Ben Carterette
University of Delaware,
Newark, DE, USA
carteret@cis.udel.edu

Paul Clough
University of Sheffield,
Sheffield, UK
p.d.clough@sheffield.ac.uk

Mark Hall
Edge Hill University, Ormskirk,
UK
hallmark@edgehill.ac.uk

Evangelos Kanoulas
University of Amsterdam,
Amsterdam, The Netherlands
e.kanoulas@uva.nl

Mark Sanderson
RMIT University, Melbourne,
Australia
mark.sanderson@rmit.edu.au

ABSTRACT

Information Retrieval (IR) research has traditionally focused on serving the best results for a single query— so-called ad hoc retrieval. However, users typically search iteratively, refining and reformulating their queries during a session. A key challenge in the study of this interaction is the creation of suitable evaluation resources to assess the effectiveness of IR systems over sessions. This paper describes the TREC Session Track, which ran from 2010 through to 2014, which focussed on forming test collections that included various forms of implicit feedback. We describe the test collections; a brief analysis of the differences between datasets over the years; and the evaluation results that demonstrate that the use of user session data significantly improved effectiveness.

1. INTRODUCTION

One of the commonest IR system evaluation methodologies is the Cranfield approach [4] using test collections to conduct controlled, systematic, and repeatable evaluations [5]. The focus of such evaluation is on how well an IR system can locate and rank relevant documents from a single query. In practice, however, users typically reformulate queries in response to search results or as their information need alters over time [7]. Retrieval evaluation should compute system success over multiple query-response interactions [8].

The TREC Session Track[1] was an attempt to evaluate IR systems over multi-query sessions. In 2010, the track produced test collections and evaluation measures for studying retrieval over sessions [9]; from 2011 on [10, 11, 1, 2], the track focused more on providing participants with user data with which to improve retrieval. The resulting collections consist of document collections, topics, and relevance assessments, as well as log data from user sessions.

[1]http://ir.cis.udel.edu/sessions/

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2914675

The track's test collections are described here and compared: including studying the effects of the search engines used to build the collections, user variability, and topic analysis. Participant results indicate that certain types of search benefit significantly from exploiting session information.

2. SESSION TRACK OVERVIEW

The aim of the track was to test if the retrieval effectiveness of a query could be improved by using previous queries, ranked results, and user interactions. We constructed four test collections comprising N sessions of varying length, each the result of a user attempting to satisfy one of T pre-defined topics. Each session numbered $1..i..N$ consisted of:

- m_i blocks of user interactions (the session's *length*);
- the *current query* q_{m_i} in the session;
- $m_i - 1$ blocks of interactions in the session prior to the current query, composed of:
  1. the user queries in the session, q_1, q_2, ..., q_{m_i-1};
  2. the ranked list of URLs seen by the user for each of those queries;
  3. the set of clicked URLs/snippets.

Ranking algorithms were evaluated on the current query under two conditions: A one-off ad hoc query; or a query using some or all of the prior logged data. The latter condition had several different sub-conditions that varied year to year: ("RL" refers to Ranked List):

- RL1: The baseline condition: an ad hoc query
- RL2-1: RL1 plus previous session queries
- RL2-2: RL2-1 plus rankings (URLs, titles, snippets)
- RL2-3: RL2-2 plus user data (clicks, dwell times)
- RL3: Using all data in the session log (in particular, other sessions on the same topic)

The focus of the track was on the degree to which a group improved their retrieval system's baseline effectiveness (RL1) by incorporating some or all of the additional log data.

3. TEST COLLECTIONS

Table 1 shows statistics of the Session track collections. The ClueWeb09 collection was used in 2011 and 2012, and the ClueWeb12 collection in 2013 and 2014.

Topics: While not a part of a true log of user search activity, we felt it was important to define topic descriptions for overall sessions so as to make relevance assessing

Table 1: Four years of TREC Session Track test collections and evaluations

	2011	2012	2013	2014
collection	ClueWeb09	ClueWeb09	ClueWeb12	ClueWeb12
topic properties				
topic set size	62	48	61	60‡
topic cat. dist.	known-item†	10 exploratory, 6 interpretive, 20 known-item, 12 known-subj	10 exploratory, 9 interpretive, 32 known-item, 10 known-subj	15 exploratory, 15 interpretive, 15 known-item, 15 known-subj
session properties				
user population	U. Sheffield	U. Sheffield	U. Sheffield + IR researchers	MTurk
search engine	BOSS+CW09 filter	BOSS+CW09 filter	indri	indri
total sessions	76	98	133	1,257
sessions per topic	1.2	2.0	2.2	21.0
mean length (in queries)	3.7	3.0	3.7	3.7
median time between queries	68.5s	66.7s	72.2s	25.6s
relevance judgments				
topics judged	62	48	49	51
total judgments	19,413	17,861	13,132	16,949
evaluation by nDCG@10				
mean RL1	0.3015	0.1847	0.1373	0.1719
mean RL2-1	0.3083	0.1950	–§	–§
mean RL2-2	0.2941	0.2140	–§	–§
mean RL2-3	0.3077	0.2303	0.1832	0.1885
mean RL3	–§	–§	0.1834	0.2002
max RL* - RL1	0.1800	0.1770	0.1230	0.1507

† 2011 topics were not categorized, but a retrospective analysis suggests most of them fit the "known-item" label best.
‡ 2014 topics were reused 2012 and 2013 topics.
§ The RL2-1 and RL2-2 conditions were eliminated for 2013 and 2014; the RL3 condition was introduced in 2013.

simpler. The challenge was to construct topics that were likely to require multiple query reformulations. In 2011, we did this by adapting multi-faceted TREC 2007 Question Answering track topics. Because of the nature of the QA track, many topics modelled "fact-finding" tasks answerable by a single document. In 2012-2013, we developed topics according to a task categorization scheme [13] with four classes: *known-item*; *known-subject*; *interpretive*; and *exploratory*. In 2014, we reused topics from 2012-2013 selecting fifteen topics from the four categories, biasing selection to topics that had longer user sessions and more clicks.

Sessions: Assessing the impact of session data on retrieval effectiveness required capturing user-system interactions, including queries, rankings, and clicks. We describe the users and search engines employed to generate the data.

Users: In 2011-2013, the primary user group were staff and students at the University of Sheffield. Using a university-wide email, we invited participants to search on as many topics as they had time for. In 2013 we solicited additional participants from the Session Track and SIG-IRList mailing lists. In 2014 we used a crowdsourcing platform (Mechanical Turk) taking a similar approach to past work for crowdsourcing interactions [15].

Search process: Users were shown a topic description, a search box for entering queries, and a list of ten ranked results with a pagination control to navigate to further results. Each retrieved item was represented by its title, URL, and snippet. Additionally, there was a "Save" button that users were instructed to use to collect those documents that helped them satisfy their information need. We experimented with additional components, such as a list of queries issued, but did not observe a difference in users' behaviour.

Search engine: In 2011-2012 we used Yahoo!'s BOSS (Build your Own Search System) API to search the live web. We filtered URLs returned by BOSS against those in the ClueWeb09 collection so that users would only see pages that were present in the publicly-available corpus. A large number of pages returned by BOSS did not match any URL in ClueWeb09. In 2013-2014, we switched to indri search with a home-built index of ClueWeb12. The indri index included each of the twenty ClueWeb12 segments (ClueWeb12_00 through ClueWeb12_19) indexed using the Krovetz stemmer and no stopword list. The indexes searched contained only text from title fields, anchor text from incoming links ("inlink" text), and page URLs. Each query was incorporated into an indri structured query language template and a retrieval score was computed from a query-likelihood model for the full document representation and three weighted combinations of 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.

The system logged all interactions with the user, including the queries issued, which documents were ranked (including URL, title, and snippet), which documents the user viewed, and which they saved as relevant to the task (note however that the latter are not the relevance judgments). This log data was then used to create the sessions.

4. EVALUATION

We used topical relevance judgments in order to compute measures of effectiveness like nDCG@10 for each topic. Since the Session Track examines whether session log data can be exploited, the evaluation examined the *change* in effectiveness from the baseline (RL1) to using some data (RL2) to using a full query log (RL3). In addition, since each topic may be the subject of more than one session, and

2013 \ 2014	4	3	2	1	0	-2
nav - 4	1	0	0	0	0	0
key - 3	0	1	2	7	4	0
hi - 2	0	4	28	52	14	2
rel - 1	1	12	75	89	64	0
not - 0	4	5	50	161	337	11
junk -2	0	0	0	0	4	5

Table 2: Agreement on relevance grades

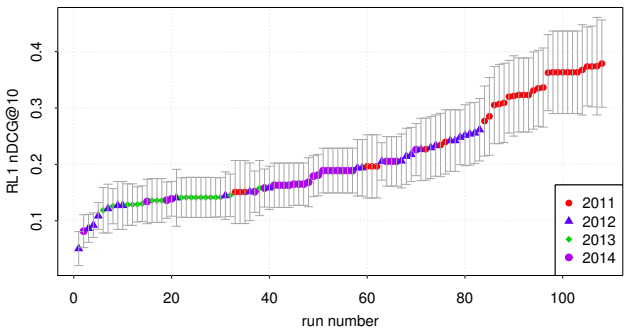

Figure 1: Mean nDCG@10 (with error bars showing ±2 standard error) for all 108 submitted runs' RL1 baseline.

each session may use different queries, the evaluation was over *sessions* rather than over topics.

Documents were selected for judging by pooling the top-10 results from all the submitted RLs along with all documents that were retrieved and viewed by the users. TREC NIST assessors (*not* the original users) judged each pooled document with respect to the topic description. All original user actions were invisible to the assessors; judgments were made solely on the topical similarity to the topic description on a 6-grade scale. Over four years, 66,548 relevance judgments were made to 60,500 unique pages identified by URL: 33,686 pages from ClueWeb09 ; 26,814 from ClueWeb12. A total of 19,179 (29%) documents were judged relevant (grade 1 or higher) and 47,369 (71%) judged nonrelevant.

Since the topics for 2014 were taken from the 2012 & 2013 Session Tracks and in the last two years the document collection was ClueWeb12, we have documents with multiple assessments. Table 2 shows assessor agreement. Assessors were much more likely to say a document judged non-relevant in 2013 was relevant in 2014 than vice versa.

Results: Figure 1 shows nDCG@10 for all groups' baseline RL1 submissions, sorted by nDCG@10 and coded by year. It is evident that 2011 had the best baseline effectiveness (average nDCG@10 of 0.30), followed by 2012 (0.18), then 2014 (0.17), and finally 2013 (0.14) had the lowest baseline effectiveness. The change from 2011 to 2012 reflects a shift to more difficult topics: the 2012 known-subject and interpretive topic categories proved to be significantly more difficult than the 2011 known-item topics. The change from 2012 to 2013 reflects a change in the underlying search technology from Yahoo! BOSS to the Indri-based system.

Figure 2 shows the improvement over each submitted run's RL1 baseline sorted by that improvement. Improvement from the RL1 baseline does not show any trend by year—for 2011, the average improvement was 0.04, for 2012 it was 0.05, for 2013 it was 0.05, and for 2014 it dropped to 0.02.

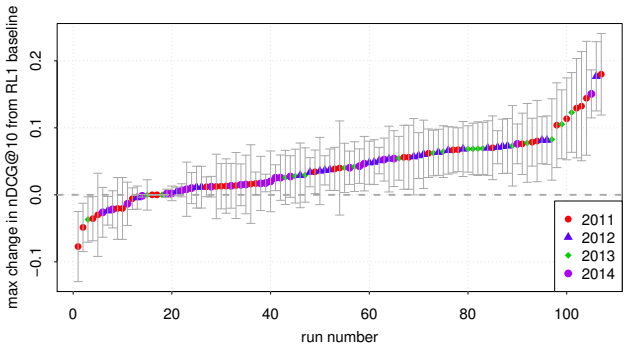

Figure 2: Largest measured improvement in nDCG@10 from RL1 to any other condition for all 108 submitted runs, with error bars showing ±2 standard errors.

From these results, we conclude that it is possible to use session history to improve effectiveness over basic ad hoc retrieval, and moreover that it does not take a lot of session history to do so. Further evidence is offered in [**?**, **?**, **?**]. A study of particular interest due to the fact that it was conducted both over a Session track collection and a commercial search engine proprietary collection is that by Raman et al. [14]; the session collection enabled them to demonstrate the effectiveness of their algorithm in accordance to the proprietary test collection.

5. ANALYSIS

In this section we perform some basic analysis of the Session Track collections and evaluation results.

Topic categories: We investigated the degree to which systems were able to improve effectiveness for each of our four topic classes. We look at the average improvement from the RL1 baseline, and find the maximum average improvement to any other RL condition for each run.

The overall mean improvements are 0.04, 0.07, 0.04, and 0.05 for known-item, known-subject, exploratory, and interpretive respectively, though only the differences between known-subject and the others were statistically significant.This suggests that known-subject topics benefit most from access to session history, but the details are more subtle. Exploratory topics tend to have the largest improvements for individual systems: the five largest improvements in exploratory topics are 5–10% larger than the five largest in known-subject topics. Exploratory topics also show the greatest benefit from the use of more log data: from RL1 to RL2, exploratory topics only increase an average of 0.03 (compare to 0.05 for known-subject topics, the largest improvement), but from RL2 to RL3 they increase by 0.05 (compare to 0.04 for known-subject topics, the second-largest improvement).

Topic variability: Most IR test collections have only one instantiation of a topic (an exception is the TREC Query track). Since we may have multiple sessions for any given topic, the Session Track gives us a chance to analyze variability in effectiveness within topics.

Figure 3 shows how much effectiveness varies over the different user sessions of a single topic. Each plot on the x-axis is a topic, the y-axis is a boxplot of the range in nDCG@10 changes from RL1 to any other RL. A taller box means more variability. A point or box plotted further up the y-axis indicates higher average change in nDCG@10 across the sessions

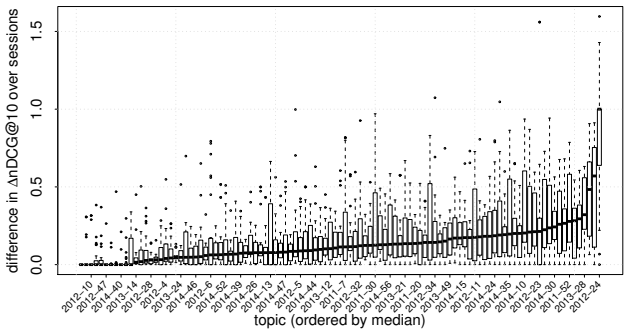

Figure 3: Variability over sessions and system effectiveness for selected topics.

of a topic. An extreme case is topic 24 from 2012 (the right most topic on the plot). There were five sessions recorded for this topic (numbers 48–52); one group improved from 0.00 in RL1 to 0.91 in RL3 on one session, but fell from 0.81 in RL1 to 0.00 in RL2 on another. Many other groups had similarly large differences across sessions on this topic.

The result indiactes that there is a substantial variability in topics, separate from the variability in system effectiveness, due to the way the users performs their search and formulates their query. Previous user study showed this as well [12]. It may be beneficial to include multiple versions of the same topic in standard test collections, so as to better capture interactions between topic and system variability.

6. CONCLUSION

This paper describes the four test collections produced for the TREC Session Track that have been used to assess the use of implicit feedback on retrieval performance within sessions. The key result from the track is that aggregate data from all participant submissions shows that retrieval effectiveness was improved for ad hoc retrieval using data based on session history data. It also appears that the more detailed the session data, the greater the improvement.

Through analyzing aspects of the test collections, such as topic categories and variability, we demonstrate how the resources can be used to investigate implicit feedback and offer reusable and publicly-accessible resources for evaluating IR systems across sessions.

7. ACKNOWLEDGEMENTS

This work was supported in part by the Australian Research Council's Discovery Projects scheme (DP130104007), the National Science Foundation (NSF) under grant number IIS-1350799, and the Google Faculty Research Award scheme. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

8. REFERENCES

[1] B. Carterette, E. Kanoulas, A. Bah, M. Hall, and P. D. Clough. Overview of the TREC 2013 Session track. In *Proceedings of TREC*, 2013.

[2] B. Carterette, E. Kanoulas, A. Bah, M. Hall, and P. D. Clough. Overview of the TREC 2014 Session track (notebook version). In *Proceedings of TREC*, 2014.

[3] 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.

[4] C. W. Cleverdon. The significance of the cranfield tests on index languages. In *Proceedings of SIGIR*, pages 3–12, 1991.

[5] D. Harman. *Information Retrieval Evaluation*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2011.

[6] 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.

[7] B. J. Jansen, D. L. Booth, and A. Spink. Patterns of query reformulation during Web searching. *J. Am. Soc. Inf. Sci. Technol.*, 60(7):1358–1371, July 2009.

[8] K. Järvelin. Explaining user performance in information retrieval: Challenges to ir evaluation. In *Proceedings of ICTIR*, pages 289–296, 2009.

[9] E. Kanoulas, B. Carterette, P. Clough, and M. Sanderson. Session track overview. In *Proceedings of the 19th Text REtreival Conference (TREC)*, 2010.

[10] E. Kanoulas, B. Carterette, M. Hall, P. D. Clough, and M. Sanderson. Overview of the TREC 2011 Session track. In *Proceedings of TREC*, 2011.

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

[12] K. S. Kim. Information-seeking on the web: Effects of user and task variables. *Library and Information Science Research*, 23(3), 2011.

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

[14] K. Raman, P. N. Bennett, and K. Collins-Thompson. Toward whole-session relevance: Exploring intrinsic diversity in web search. In *Proceedings of SIGIR*, pages 463–472, 2013.

[15] G. Zuccon, T. Leelanupab, S. Whiting, E. Yilmaz, J. Jose, and L. Azzopardi. Crowdsourcing interactions: Capturing query sessions through crowdsourcing. In Carterette et al. [3].