CLEF 2017 Dynamic Search Lab Overview And Evaluation

Evangelos Kanoulas[1] and Leif Azzopardi[2]

[1] Informatics Institute, University of Amsterdam, Netherlands, E.Kanoulas@uva.nl
[2] Computer and Information Sciences, University of Strathclyde, UK, leif.azzopardi@strath.ac.uk

Abstract. In this paper we provide an overview of the first edition of the CLEF Dynamic Search Lab. The CLEF Dynamic Search lab ran in the form of a workshop with the goal of approaching one key question: how can we evaluate dynamic search algorithms? Unlike static search algorithms, which essentially consider user request's independently, and which do not adapt the ranking w.r.t the user's sequence of interactions, dynamic search algorithms try to infer the user's intentions from their interactions and then adapt the ranking accordingly. Personalized session search, contextual search, and dialog systems often adopt such algorithms. This lab provides an opportunity for researchers to discuss the challenges faced when trying to measure and evaluate the performance of dynamic search algorithms, given the context of available corpora, simulations methods, and current evaluation metrics. To seed the discussion, a pilot task was run with the goal of producing search agents that could simulate the process of a user, interacting with a search system over the course of a search session. Herein, we describe the overall objectives of the CLEF 2017 Dynamic Search Lab, the resources created for the pilot task, the evaluation methodology adopted, and some preliminary evaluation results of the Pilot task.

1 Introduction

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. IR systems can still respond to each query in a session independently of the history of user interactions, or alternatively adopt their model of relevance in the context of these interactions. A key challenge in the study of algorithms and models that dynamically adapt their response to a user's query on the basis of prior interactions is the creation of suitable evaluation resources and the definition of suitable evaluation metrics to assess the effectiveness of such IR algorithms. Over the years various initiatives have been proposed which have tried to make progress on this long standing challenge.

The TREC Interactive track [8], which ran between 1994 and 2002, investigated the evaluation of interactive IR systems and resulted in an early standardization of the experimental design. However, it did not lead to a reusable test

collection methodology. The High Accuracy Retrieval of Documents (HARD) track [1] followed the Interactive track, with the primary focus on single-cycle user-system interactions. These interactions were embodied in clarification forms which could be used by retrieval algorithms to elicit feedback from assessors. The track attempted to further standardize the retrieval of interactive algorithms, however it also did not lead to a reusable collection that supports adaptive and dynamic search algorithms. The TREC Session Track [3], which ran from 2010 through 2014, made some headway in this direction. The track produced test collections, where included with the topic description was the history of user interactions with a system, that could be used to improve the performance of a given query. While, this mean adaptive and dynamic algorithms could be evaluated for one iteration of the search process, the collection's are not suitable for assessing the quality of retrieval over an entire session. In 2015, the TREC Tasks track [14,13], a different direction was taken, where the test collection provides queries for which all possible sub-tasks did to be inferred, and the documents relevant to those sub-tasks identified. Even though the produced test collections could be used in testing whether a system can help the user to perform a task end-to-end, the focus was not on adapting and learning from the user's interactions as in the case of dynamic search algorithms.

In the related domain of dialogue systems, the advancement of deep learning methods has led to a new generation of data-driven dialog systems. Broadly-speaking, dialog systems can be categorized along two dimensions, (a) goal-driven vs. non-goal-driven, and (b) open-domain vs. closed domain dialog systems. Goal-driven open-domain dialog systems are in par with dynamic search engines: as they seek to provide assistance, advice and answers to a user over unrestricted and diverse topics, helping them complete their task, by taking into account the conversation history. While, a variety of corpora is available for training such dialog systems [12], when it comes to the evaluation, the existing corpora are inappropriate. This is because they only contain a static set of dialogues and any dialog that does not develop in a way similar to the static set cannot be evaluated. Often, the evaluation of goal-driven dialogue systems focuses on goal-related performance criteria, such as goal completion rate, dialogue length, and user satisfaction. Automatically determining whether a task has been solved however is an open problem, while task-completion is not the only quality criterion of interest in the development of dialog systems. Thus, simulated data is often generated by a simulated user [6,4,11]. Given a sufficiently accurate model of how user's converse, the interaction between the dialog system and the user can be simulated over a large space of possible topics. Using such data, it is then possible to deduce the desired metrics. This suggests that a similar approach could be taken in the context of interactive IR. However, while significant effort has been made to render the simulated data as realistic as possible [7,9], generating realistic user simulation models remains an open problem.

2 Lab Overview

The CLEF Dynamic Tasks Lab attempts to focus attention towards building a bridge between batch TREC-style evaluation methodology and the Interactive Information Retrieval evaluation methodology - so that dynamic search algorithms can be evaluated using re-usable test collections.

The objectives of the lab is threefold:

1. to devise a methodology for evaluating dynamic search algorithms by exploring the role of simulation as a means to create re-usable test collections,
2. to develop evaluation metrics that measure the quality during the session, both at different stages of the search process and at the end of the session,
3. to develop algorithms that can provide an optimal response in an interactive retrieval setup.

The focus of the CLEF 2017 Dynamic Tasks Lab is to provide a forum that can help foster research around the evaluation of dynamic retrieval algorithms. The Lab, in the form of a Workshop, solicits the submission of two types of papers: (a) position papers, and (b) data papers. The goal of position papers was evaluation methodologies for assessing the quality of search algorithms with the user in the loop, under two constraints: any evaluation framework proposed should allow the (statistical) reproduciblility of results, and lead to a reusable benchmark collection. The goal of the data papers was to describe test collections or data sets suitable for guiding the construction of dynamic test collections, tasks and evaluation metrics.

3 Pilot Task

Towards the aforementioned goals of generating simulation data the CLEF 2017 Dynamic Tasks Lab ran a pilot task in the context of developing Task Completion Engines [2] and Intelligent Search Agents [7]. Task Completion Engines and Autonomous Search Agents are being developed to help users in acquire information in order to make a decision and complete a search task. At the same time such Intelligent Search Agents, encode a model of a user, and so present the potential to simulate users submitting queries, which can enable the evaluation of dynamic search algorithms. Such engines/agents need to work with a user to ascertain their information needs, then perform their own searches to dynamically identify relevant material, which will be useful in completing a particular task. For example, consider the task of organizing a wedding. There are many different things that need to be arranged and ordered, e.g. a venue, flowers, catering, gift list, dresses, car hire, hotels, etc. Finding relevant sites and resources requires numerous searches and filtering through many documents/sites. A search agent could help to expedite the process by finding the relevant sites to visit, while a task completion engine would provide a structured interface to help complete the process.

In this year's Dynamic Search Task Track, the task can be interpreted in one of two ways:

1. to generate a series of queries that a search agent would issue to a search engine, in order to compile a set of links useful for the user. This set might be presented to the user or used for further processing by a task completion engine; or
2. to generate a series of query suggestions that a search engine would recommend to the user, and thus the suggested course of interaction.

As a starting point, for building a test collection, we first consider how people look for information required to complete various casual leisure and work tasks. The history of queries and interactions are then used as a reference point during the evaluation to determine if agents/dynamic search algorithms/query suggestion algorithms that can generate queries that are like those posed by people. And thus, see how well human like queries can be generated for suggestions)? Thus the focus of the track is on query generation and models of querying, based on task and interaction history.

3.1 Task Description and Data Sets

Starting with an initial query for a given topic, the task is to generate a series of subsequent or related queries. The data used is TREC ClueWeb Part B test collection and the topics used are sourced from the Session Track 2014 [3]. A fixed search engine was setup, where ClueWeb was indexed using ElasticSearch. The title, url, page contents were indexed, along with the spam rank and page rank of each document. The ElasticSearch API was then provided as the "search engine" that the agent or person is using to undertake each task/topic. From the Session Track 2014 topics, a subset of 26 topics were selected out of the original 50, based on the following criteria: there were four or more queries associated with the topic, where the subsequent interaction on each query lead to identifying at least one TREC relevant document. These were considered, good or useful, queries i.e. they helped identify relevant material. The set of "good" queries were with-held as the relevant set. The TREC topic title, was provided to participants as the initial seed query i.e. the first query in the session.

Interaction data was then provided to provide simulated interaction with queries issued to the search engine. It was anticipated that the simulated clicks could be used by the algorithms to help infer relevance of the documents. This data could be used could be used as (a) a classifier providing relevance decisions regarding observed items in the result list, or (b) as clicks that the user performed when viewing the results of a query (i.e. given a query, assume that this is what the user clicks on, to help infer the next query). A set of judgments/click was generated based on the probability of Session Track users clicks data, conditioned by relevance (i.e. the probability of a click, if then document was TREC Relevant or TREC Non-Relevant).

The task, then, was to provide a list of query suggestions/recommendations along with a list of up to 50 documents.

3.2 Participants

Two teams participated in this Pilot Task: (1) the Web Technology and Information Systems, Bauhaus-Universitat Weimar, Germany (Webis), and (2) the Vienna University of Technology (TUW). The Webis team submitted on set of query suggestions, and a list of up to 50 documents. The TUW team focused on the document retrieval, submitting only a ranked list of documents.

Webis [5] evaluated query suggestions in form of keyqueries for clicked documents. A query q is a keyquery for a document set D iff q returns the documents from D in its top-k ranks, q has at least l results, and no subquery of q has the previous two properties. Our query suggestion approach derives keyqueries for pairs of documents previously clicked by the user. The Dynamic Search Lab data contains 26 topics, along with one query submitted by some user in a search session and the shown results from the whole session with some indicated as being clicked by the user. Exactly for these clicked documents, they derive keyqueries as query suggestions. Regarding the ranked list for a topic, they used the top-10 results of each of the at most five derived keyqueries returned by the Dynamic Search API and merged them as follows: first the first ranks of the queries, then the second ranks, etc.; duplicate results that already were in the merged list were replaced by the next result from the same keyquery. Note that the Webis team submitted results for 19 out of the 26 topics.

TUW [10] propose the creation of a search agent that specifically leverages the structure of Wikipedia articles to understand search tasks. Their assumption is that human editors carefully choose meaningful section titles to cover the various aspects of an article. Their proposed search agent is responsible for two tasks: identifying the key Wikipedia articles related to a complex search task, and selecting section titles from those articles. TUW contributed 5 runs: TUW_0_baseline.run is an ElasticSearch BM25 run that uses the query field for each information need provided; TUW_1 human.run is a manual run where a human judge selected up to 5 section titles from the top Wikipedia articles returned by the Wikipedia API; TUW_2_First_Five.run automatically choosing the first five sections of Wikipedia, TUW_3_Word2VecMean.run used as background text the description of the information need provided and compared it to the text used in each individual section of Wikipedia; it made use of the cosine similarity between the vector representation of each word in the both texts, with text being represented by Word2Vec trained on GoogleNews; TUW_4_W2V_Plus_NB.run extended the TUW_3 run by predicting the relevance of each retrieved document automatically before applying the round-robin algorithm to merge the result of each query.

3.3 Evaluation

Given that evaluation is an open problem, the lab was also open to different ways in which to evaluate this task. Some basic measures that were considered are:

– Query term overlap: how well do the query terms in the suggestions match with the terms used
– Query likelihood: how likely are the queries suggested given the model of relevance.
– Precision and Recall based measures on the set of documents retrieved.
– Suggest your own measure: how to measure the usefulness and value of queries is a rather open question. So how can we evaluate how good a set of queries are in relation to completing a particular task?

During the lab, we will discuss the various challenges of constructing reusable test collections for evaluating dynamic search algorithms and how we can develop appropriate evaluation measures.

3.4 Results

The intrinsic evaluation of the suggested queries is a hard problem. Therefore the preliminary results for the Pilot task focus on the extrinsic evaluation of these queries, by evaluating the ranked list of retrieved document. Table 1 contains mean values of precision at cut-off 50, recall at cut-off 50, and average precision. Figure 1 presentes the box plots for average precision across the 26 topics for TUV and 19 topics for Webis.

Run	P@50	R@50	AP
tuw0_only_query	0.1177	**0.0504**	0.0181
tuw1_human	0.1031	0.0472	0.0147
tuw2_fifo	0.0885	0.0393	0.0099
tuw3_mean	0.0900	0.0409	0.0133
tuw4_nb_mean	0.0254	0.0076	0.0020
webis-result-list	**0.1411**	0.0502	**0.0194**

Table 1. Evaluation of the submitted ranked lists of results by participants.

4 Conclusions

The goal of the CLEF 2017 Dynamic Search lab is to answer one key question: how can we evaluate dynamic search algorithms? Towards this question one position paper was submitted, and will be presented during the CLEF workshop. Further, a Pilot task was put in place, using data from the TREC 2014 Session track, in an attempt to test user simulations. Two teams participated in the task, submitting a total of 6 runs. One team submitted suggested queries, while both teams submitted ranked lists of results. The intrinsic evaluation of queries appeared to be a hard problem, hence participants were evaluated extrinsically, on the basis of their submitted ranked lists. The results are preliminary to draw any conclusions.

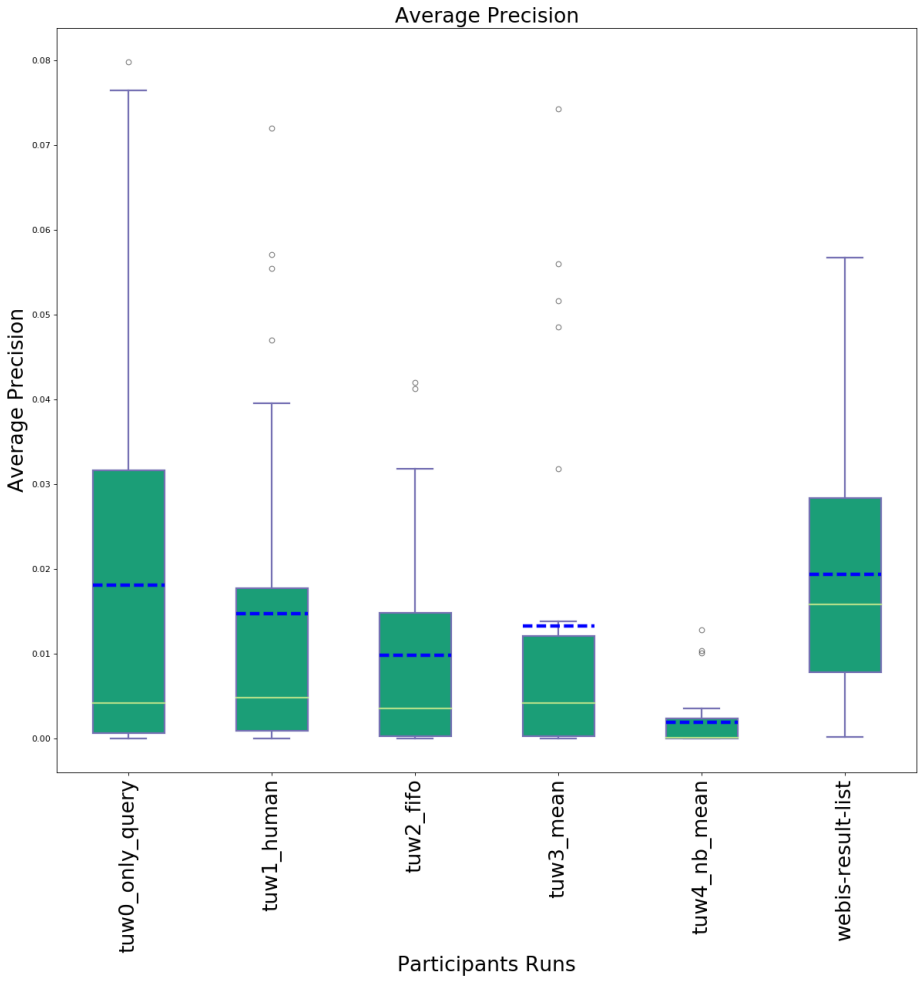

Fig. 1. Box-plots of average precision values across the topics; the blue dashed line represents the mean average precision.

Acknowledgements

This work was partially supported by the Google Faculty Research Award program and the Microsoft Azure for Research Award program (CRM:0518163). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors. We would also like to thank Dr. Guido Zuccon for setting up the ElasticSearch API.

References

1. Allan, J.: Hard track overview in trec 2003 high accuracy retrieval from documents. Tech. rep., DTIC Document (2005)
2. Balog, K.: Task-completion engines: A vision with a plan. In: SCST@ECIR (2015)
3. Carterette, B., Clough, P.D., Hall, M.M., Kanoulas, E., Sanderson, M.: Evaluating retrieval over sessions: The TREC session track 2011-2014. In: Perego, R., Sebastiani, F., Aslam, J.A., Ruthven, I., Zobel, J. (eds.) Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016. pp. 685–688. ACM (2016), http://doi.acm.org/10.1145/2911451.2914675
4. Georgila, K., Henderson, J., Lemon, O.: User simulation for spoken dialogue systems: learning and evaluation. In: Interspeech. pp. 1065–1068 (2006)
5. Hagen, M., Kiesel, J., Alshomary, M., Stein, B.: Webis at the clef 2017 dynamic search lab. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)
6. Jung, S., Lee, C., Kim, K., Jeong, M., Lee, G.G.: Data-driven user simulation for automated evaluation of spoken dialog systems. Comput. Speech Lang. 23(4), 479–509 (Oct 2009), http://dx.doi.org/10.1016/j.csl.2009.03.002
7. Maxwell, D., Azzopardi, L.: Agents, simulated users and humans: An analysis of performance and behaviour. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 731–740. CIKM '16 (2016)
8. Over, P.: The trec interactive track: an annotated bibliography. Information Processing & Management 37(3), 369–381 (2001)
9. Pääkkönen, T., Kekäläinen, J., Keskustalo, H., Azzopardi, L., Maxwell, D., Järvelin, K.: Validating simulated interaction for retrieval evaluation. Information Retrieval Journal pp. 1–25 (2017)
10. Palotti, J.: Leveraging wikipediaâĂŹs article structure to build search agents: Tuw at clef 2017 dynamic search. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)
11. Pietquin, O., Hastie, H.: A survey on metrics for the evaluation of user simulations. The knowledge engineering review 28(01), 59–73 (2013)
12. Serban, I.V., Lowe, R., Henderson, P., Charlin, L., Pineau, J.: A survey of available corpora for building data-driven dialogue systems. CoRR abs/1512.05742 (2015), http://arxiv.org/abs/1512.05742
13. Verma, M., Yilmaz, E., Mehrotra, R., Kanoulas, E., Carterette, B., Craswell, N., Bailey, P.: Overview of the TREC tasks track 2016. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016,

Gaithersburg, Maryland, USA, November 15-18, 2016. vol. Special Publication 500-321. National Institute of Standards and Technology (NIST) (2016), http://trec.nist.gov/pubs/trec25/papers/Overview-T.pdf

14. Yilmaz, E., Verma, M., Mehrotra, R., Kanoulas, E., Carterette, B., Craswell, N.: Overview of the TREC 2015 tasks track. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015. vol. Special Publication 500-319. National Institute of Standards and Technology (NIST) (2015), http://trec.nist.gov/pubs/trec24/papers/Overview-T.pdf