Overview of TREC OpenSearch 2017

Rolf Jagerman
University of Amsterdam
Amsterdam, The Netherlands
rolf.jagerman@uva.nl

Krisztian Balog
University of Stavenger
Stavenger, Norway
krisztian.balog@uis.no

Phillip Schaer
TH Köln (University of Applied Sciences)
Köln, Germany
phillip.schaer@th-koeln.de

Johann Schaible
GESIS — Leibniz Institute for the Social Sciences
Köln, Germany
johann.schaible@gesis.org

Narges Tavakolpoursaleh
GESIS — Leibniz Institute for the Social Sciences
Köln, Germany
narges.tavakolpoursaleh@gesis.org

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
derijke@uva.nl

ABSTRACT

In this paper we provide an overview of the TREC 2017 OpenSearch track. The OpenSearch track provides researchers the opportunity to have their retrieval approaches evaluated in a live setting with real users. We focus on the academic search domain with the Social Science Open Access Repository (SSOAR) search engine and report our results.

1 INTRODUCTION

The goal of Information Retrieval (IR) is to help people find information. Experiments on IR methods conducted in the setting of community-based benchmarking efforts have traditionally been done using either datasets made by professional assessors or using simulated users. This type of set up helps the repeatability and reproducibility of experiments. A drawback, however, is that these types of experiments largely abstract away the user. Furthermore, there are several limitations: (1) obtaining relevance judgements by professional assessors does not scale and is expensive, (2) relevance may change over time and is not a static concept, and (3) relevance judgements may not accurately portray the intents of the real users. A way to over-come these limitations is to use *online evaluation*, where we observe users *in situ* and measure metrics such as click-through rate, time-to-success, abandonment, etc.

Unfortunately, access to real users is reserved for owners of online properties with a large and active user-base. There are considerable engineering and logistic challenges in setting up a search service and attracting a large user-base. As a result, there is a gap between the evaluation methods of researchers who have access to real users and those who do not. The aim of TREC OpenSearch is to bridge this gap and make evaluation using real users open to all researchers:

> "Open Search is a new evaluation paradigm for IR. The experimentation platform is an existing search engine. Researchers have the opportunity to replace components of this search engine and evaluate these components using interactions with real, unsuspecting users of this search engine" [2].

In this paper we give a brief overview of the results of the OpenSearch track during 2017. We first provide a description of the living labs methodology in Section 2. Next, we discuss the academic search use-case at SSOAR in Section 3. We present the outcomes of the track in Section 4. Finally, we conclude the overview in Section 5.

2 LIVING LABS METHODOLOGY

We consider the living labs methodology in a search engine setting: Users submit queries and obtain a ranked list of documents. An overview of the living-labs setup that we use is displayed in Figure 1 and will now be described in more detail.

A real search engine extracts a set of queries $\mathbb{Q}$ from their logs. The queries are chosen in such a way that it is likely that they will be issued again in the future. For each query q there is also a corresponding set of candidate documents $\mathbb{D}_q$. The sites submit this information to the living-labs API[1] [1] (step 1 in Figure 1), after which it can be downloaded by the participants of the track (step 2 in Figure 1).

Participants are asked to submit their ranked lists of documents for each of the queries submitted by the sites. They are free to use any ranking algorithms they deem appropriate. In essence this task boils down to an ad-hoc document retrieval task (restricted to the candidate set $\mathbb{D}_q$). The ranked lists produced by the participant's method are then submitted to the living-labs API (step 3 in Figure 1).

When a user of the site issues a query $q \in \mathbb{Q}$, the site will first obtain an experimental ranking from one of the participants via the living-labs API. This ranking is then interleaved, using the team-draft interleaving algorithm [3], with the production ranking. This produces the SERP that is shown to the user (step 4 in Figure 1). The user interacts with this SERP by clicking on some of the entries. The clicks are recorded and submitted back to the API in the form of feedback.

[1] https://bitbucket.com/living-labs/ll-api/

Proceedings of the Twenty-Sixth Text REtrieval Conference (TREC 2017). NIST,
 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

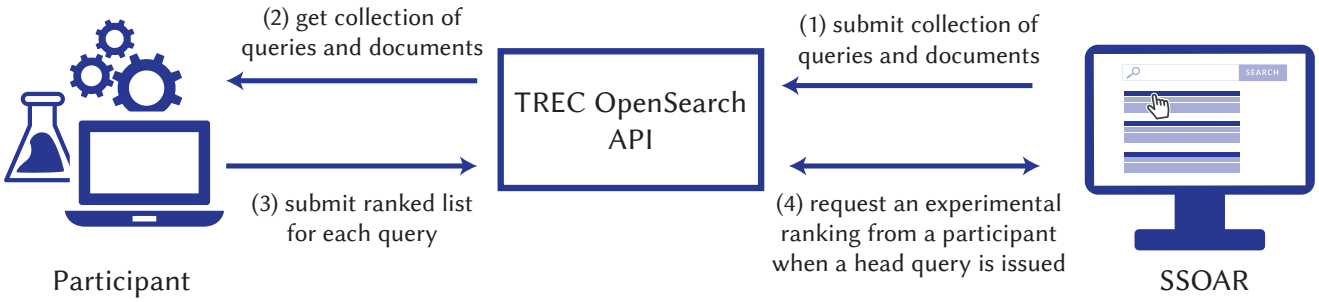

Figure 1: An overview of TREC OpenSearch. SSOAR provides their collection of queries and documents to the OpenSearch API. Participants download this collection and then compute and submit their ranked lists. Whenever a user on SSOAR issues one of the selected queries, SSOAR requests an experimental ranking from a random participant and interleaves it with their production ranking to produce the SERP.

The team-draft interleaving algorithm attributes clicks to either the production ranking or the participant's ranking. By counting the number of times a click is attributed to either one of these, we can infer a winner or a tie for each impression (i.e., observation of the result list upon issuing of a query).

2.1 Changes Since Last Edition

To improve the track we have made numerous improvements and modifications to the API code since the 2016 edition. Compared to the 2016 edition of TREC OpenSearch, there are four new major features:

2.1.1 Interleaving endpoint. The API has a new endpoint for interleaving. This endpoint allows sites to submit their production ranking and obtain an interleaved ranked list. This reduces engineering overhead for participating sites, as they will not have to implement the interleaving algorithm themselves but can instead rely on the one provided by the API.

2.1.2 Authentication. We switched to the HTTP Basic authentication scheme to authenticate communication with the API. In previous years we would authenticate users by checking the API key that was provided in the URL. The new authentication scheme follows the HTTP Basic authentication standard and is more secure.

2.1.3 Multiple Runs. Participants are now able to submit up to five runs. Impression traffic is still distributed fairly amongst all participants, i.e. having multiple runs does not give you more impressions. Although this feature was much requested in the previous edition of the track, it was not used by the participants featured in this paper.

2.1.4 Load-balancing. In previous editions, traffic was distributed uniformly at random. We removed the random component and replaced it with a load-balancer that prioritizes participants with the least amount of impressions so far. Unfortunately this load-balancer caused an issue with one participant's run. Their runs were not displayed during the competition, because they were activated very early and had gathered too many impressions in the time leading up the start of the round. During the real round the load-balancer tried to compensate for these impressions by prioritizing the other participants. This behavior was unintentional and subsequently fixed after uncovering the problem. Unfortunately the round had already concluded by this point, so we are not able to provide results to this participant.

3 ACADEMIC SEARCH AT SSOAR

The participating site for TREC OpenSearch 2017 is the Social Science Open Access Repository (SSOAR)[2], which generously provided us with a collection of queries and documents to be ranked, and opened up their service to the experimental rankings made by the participants.

SSOAR is an open access document repository that is based on the Solr-based software DSpace[3] which is one of the state-of-the-art repository systems in the open access community. SSOAR is developed and maintained at GESIS, Cologne, Germany. It contains over 43,500 full text documents from the social sciences and neighboring areas. Each document is annotated with a rich set of metadata, mostly including descriptors and classification information. Around 60,000 unique visitors visit SSOAR and download more than 213,000 PDF full texts per month. Both numbers are cleaned from search engine accesses using the enterprise web tracking software E-Tracker.

The queries and candidate documents were extracted from the live system. We decided to include more than 1000 queries (instead of 100) to allow generating more impressions and hopefully more clicks as well. In detail, we used the most frequent 1200 head queries from a one year log file dump of SSOAR. Several example queries are displayed in Table 1. After a manual filtering process that erased some obvious data gibberish, 1165 queries remained. The queries were split into 500 test and 665 training queries. The training/test split was non-uniform by mistake. In particular, we took the top frequent queries as test queries and the remaining as training queries. Unfortunately, this type of split leads to training and test queries with very different characteristics and is something that should be avoided in the future.

The items in the set of candidate documents consist of a title and additional content descriptions including some of the rich

[2]http://www.ssoar.info/
[3]http://www.dspace.org/introducing

metadata extracted from the DSpace system. We included information like abstracts, author names, publishers, language, volume and issue numbers, subjects, and if available thesaurus descriptors.[4] Additionally, the original document in SSOAR comprises a persistent identifier and the original document ID, such that even more document metadata—that is available in English as well as in German—can be extracted from the OAI-PMH interface of SSOAR.[5] An example document is displayed in Listing 1.

3.1 Changes Since Last Edition

In contrast to the 2016 edition of the OpenSearch track, we did not include queries gathered from the discovery interface of SSOAR,[6] as these turned out to produce mostly tie results. The reason behind this is that the general pattern of these discovery queries is to start the search process with a selection of a topic from a topical classification of the social sciences. Subsequently, users tend to drill down their search by using facets like publication year or authors. Such drill down searches have the effect that they are not covered by the top 100 documents precomputed by the teams participating in OpenSearch. Therefore, these queries produce a vast amount of ties that do not help to evaluate the different participating systems.

The technical infrastructure of SSOAR to provide the Living Lab functionality was expanded. In last year's OpenSearch track, we implemented the Living Lab functionality directly on the server hosting the SSOAR live system. This led to various performance issues of the live system due to communication time-outs and the performance-heavy feedback submitted back to the Living Labs API. To address these issues, we set up and configured an own server for the Living Lab functionality. The server was used to extract the head queries and candidate documents from the SSOAR log files, to communicate with the Living Labs API when a head query was triggered, as well as to record, compute, and submit the feedback to the Living Labs API. As a result, if one of these tasks were to cause any problems, the functionality of the SSOAR live system would not be affected. The rest of the technical infrastructure is the same as last year. For more technical details please check the corresponding section in [2].

Table 1: Example queries.

ID	Query string
ssoar-q43	migration
ssoar-q115	bilateral relations
ssoar-q289	labor sozialwissenschaft
ssoar-q376	brexit
ssoar-q482	gruppendynamik
ssoar-q699	alkohol
ssoar-q803	migration und gesundheit

[4] http://lod.gesis.org/thesoz/en.html
[5] http://www.ssoar.info/OAIHandler/request?verb=Identify#
[6] http://www.ssoar.info/ssoar/discover

Listing 1: Example SSOAR Document


```
{
        "docid": "ssoar-d10466",
        "content": {
                "abstract": "Plausibilit\u00e4t spielt in allen Wissenschafts...
                "author": "Reszke, Paul",
                "available": "2015-12-14T11:20:34Z",
                "description": "Published Version",
                "identifier": "urn:nbn:de:0168-ssoar-455901",
                "issued": "2015",
                "language": "de",
                "publisher": "DEU",
                "subject": "10200",
                "type": "collection article"
        },
        "creation_time": "2017-06-15T17:04:07.403+0200",
        "site_id": "ssoar",
        "title": "Linguistic-philosophical investigations of plausibility: pat...
}
```


4 RESULTS

In this section we present the results of running the TREC OpenSearch track in 2017. We report on a single evaluation round, which ran during August 1st–August 31st 2017. The main results are displayed in Table 2. For each participating system we show the number of impressions, clicks, wins, ties, losses and the outcome. Outcome is our main evaluation metric, summarizing the performance of each system against the production ranker, and is computed as follows:

$$\text{Outcome} = \frac{\#\text{Wins}}{\#\text{Wins} + \#\text{Losses}}. \quad (1)$$

Out of the participating teams, Gesis was the winning team with the highest outcome score.

Unfortunately, the number of clicks is insufficient to draw statistically significant conclusions. We perform the sign test where our null-hypothesis is that there is no preference, i.e. each system has a 50% chance to win. The p-values we found were 0.61, 0.99 and 0.61 for teams Gesis, Webis and ICTNET respectively. These numbers tell us that the observed differences are not statistically significant and we would need to collect more data. Assuming the ratio of wins and losses stays the same, we would need to collect about 9 months worth of data to declare a winner with a p-value < 0.05.

Table 2: Outcome of TREC OpenSearch 2017 for SSOAR.

	Imps	Clicks	Wins	Ties	Losses	Outcome
Gesis	3658	31	9	2	6	0.6
Webis	3662	30	6	3	7	0.462
ICTNET	3191	44	6	4	9	0.4

We plot the number of impressions distributed across queries in Figure 3. The impressions follow a power law distribution; several queries are responsible for many of the impressions, while most queries are only issued a handful of times. The distribution of clicks follows a similar distribution as is displayed in Figure 4.

We note that the click-through rate for this round was particularly low. Out of the many thousands of impressions only a few dozen resulted in a click. A further analysis of the traffic shows that some queries were issued on a regular interval. In particular we see in Figure 2 that query `ssoar-q1` is issued exactly every 5 minutes. This tells us that some automated process, such as a crawler or a bot, is requesting this query. However, since this query also

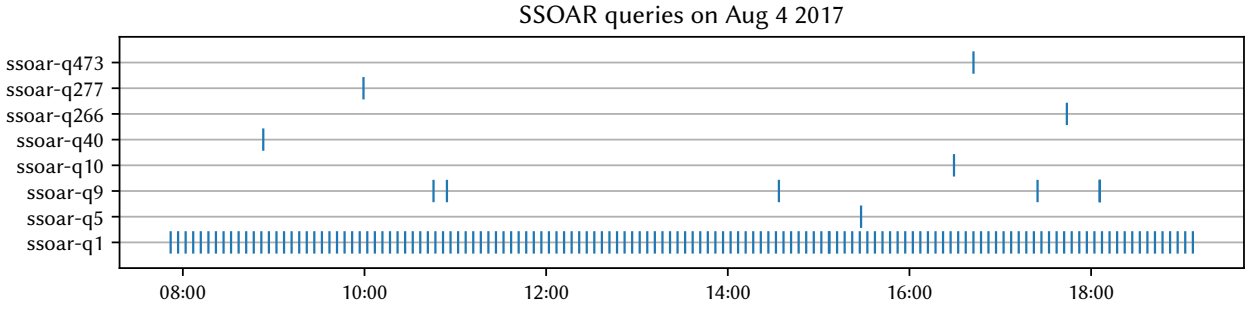

Figure 2: Query frequency over time on August 4th 2017. Each bar indicates when a query was issued. Notice that `ssoar-q1` was issued exactly every 5 minutes, indicating a crawler or bot.

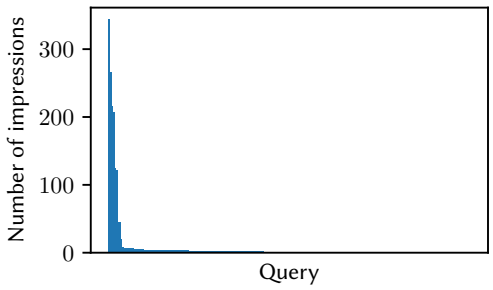

Figure 3: The number of impressions distributed over the queries. We removed `ssoar-q1` before plotting to make this plot more readable.

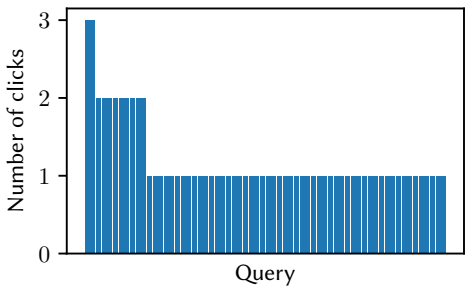

Figure 4: The number of clicks distributed over the queries.

occasionally resulted in real clicks, we cannot easily distinguish the automated traffic from the real human traffic. Naively removing this query would also remove valuable click data, so we decided to leave this data in.

We are running an extra round during October 2017 to obtain more clicks. Furthermore, this extra round offers the team who experienced a problem with the new traffic load-balancer (see Section 2.1.4) a chance to compete with the other teams. At the time of writing, this extra round is still underway.

5 CONCLUSION

In this paper we present our results from the TREC 2017 OpenSearch track. The infrastructure for the OpenSearch track was kept largely the same as the 2016 version of the track, with some minor modifications and improvements. The results show that traffic for this round was low and clicks were extremely sparse. This makes it difficult to draw statistically significant conclusions.

Online evaluation remains an important part of IR. The necessity of evaluation using real users is becoming increasingly apparent with the development of new technologies such as conversational assistants [4]. We can no longer rely solely on Cranfield-style evaluations. With our work we hope to have made the first steps towards an online evaluation platform that is open to all researchers.

Acknowledgments

We would like to thank Anne Schuth and Peter Dekker for their help in developing and setting up the infrastructure for TREC OpenSearch. Additionally we are thankful to SSOAR for sharing their queries, documents and users. We also want to thank all the participating teams. This research was supported by Ahold Delhaize, Amsterdam Data Science, the Bloomberg Research Grant program, the Criteo Faculty Research Award program, Elsevier, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the Microsoft Research Ph.D. program, the Netherlands Institute for Sound and Vision, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.001.116, HOR-11-10, CI-14-25, 652.002.001, 612.001.551, 652.001.003, and Yandex. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Krisztian Balog, Liadh Kelly, and Anne Schuth. 2014. Head First: Living Labs for Ad-hoc Search Evaluation. In *Proc. of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. 1815–1818.

[2] Krisztian Balog, Anne Schuth, Peter Dekker, Narges Tavakolpoursaleh, Philipp Schaer, Po-Yu Chuang, Jian Wu, and C. Lee Giles. 2016. Overview of the TREC 2016 Open Search track: Academic Search Edition. In *Proceedings of the Twenty-Fifth Text REtrieval Conference (TREC '16)*.

[3] Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. 2012. Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems (TOIS)* 30, 1 (2012), 6.

[4] Julia Kiseleva and Maarten de Rijke. 2017. Evaluating Personal Assistants on Mobile devices. *arXiv preprint arXiv:1706.04524* (2017).