Does Online Evaluation Correspond to Offline Evaluation in Query Auto Completion?

Alexandros Bampoulidis(✉)[1], João Palotti[1], Mihai Lupu[1], Jon Brassey[2], and Allan Hanbury[1]

[1] TU Wien, Favoriten Strasse 9-11/188, Vienna, AT
{name.surname}@tuwien.ac.at
[2] Trip Database Ltd., Little Maristow, Glasllwch Lane, Newport, UK
jon.brassey@tripdatabase.com

Abstract. Query Auto Completion is the task of suggesting queries to the users of a search engine while they are typing a query in the search box. Over the recent years there has been a renewed interest in research on improving the quality of this task. The published improvements were assessed by using offline evaluation techniques and metrics. In this paper, we provide a comparison of online and offline assessments for Query Auto Completion. We show that there is a large potential for significant bias if the raw data used in an online experiment is re-used for offline experiments afterwards to evaluate new methods.

1 Introduction

Search logs are the traces that users leave behind when searching for information with a search engine. A number of techniques benefit from analyzing them. Among those, Query Auto Completion (QAC) helps users express their information need by suggesting queries before issuing one. This work is focused on comparing the online and offline evaluation for QAC.

Query Auto Completion is the task of suggesting full queries (*completions*) to the user, which are extensions of what he/she has typed so far (*prefix*). The simplest QAC approach is Most Popular Completion (MPC) [1]. MPC ranks the completions that match the prefix by popularity. More advanced approaches are time-sensitive [3, 11] and user-sensitive [1, 6], which take into consideration the timeframe and the user's search history, behavior and profile.

Regardless of the method, a significant challenge here is the evaluation of the ranking of the completions. For practical reasons, offline evaluation is the method of choice for researchers in academia.

The central idea of offline QAC evaluation is simulating clicks on completions: Each unique query that the users have issued, as extracted from the search logs, is treated as if they clicked on it as a completion. A list of completions is generated offline, having as prefix various substrings of the query (usually 1-20 first characters) and, given the position of the query in the completions list, a score is calculated for an evaluation metric.

Mean Reciprocal Rank (MRR), a precision-oriented metric, is the standard for QAC. Other metrics are weighted MRR [1], which takes into consideration the number of completions available for the given prefix, and Success Rate at top K (SR@K) [6]. Recent studies, where QAC approaches were compared and evaluated offline, were conducted in $[2, 3, 5, 6, 10, 11]$.

However, none of these studies compare their offline results with online experiments. This is, as we will show, vital, because if we cannot control how the query logs were generated (which QAC method was used in the online system), we will observe misleading results.

To consider online evaluation, we have to draw inspiration from a higher level task: search effectiveness. We have two options: AB testing and Interleaving. AB testing is the standard of online evaluation in IR, used in predicting user satisfaction [8]. It is a controlled experiment where some of the users are exposed to an experimental version of the system. AB testing has low sensitivity due to the high variance of the users and require millions of interactions in order to reach a valid conclusion as to which system is preferred by the users [4].

Interleaving [7] exposes the users to a system which mixes an experimental version and the baseline together in as unbiased a way as possible. Interleaved comparisons have high sensitivity and require much fewer interactions than AB testing [4]. The most widely used algorithm is Team-Draft Interleaving (TDI) [9], which is the one we used in our experiments. User preference is inferred by counting the clicks credited to one version or the other (see Sect. 3).

To the best of our knowledge, no prior work has been published on online QAC evaluation and QAC interleaved comparisons. This fact raises the **research question:** Does online evaluation correspond to offline evaluation in QAC? The answer is *tentatively* no. The results are discussed in Section 4.

2 Experimental Methodology

As the core of this work is investigating the link between online and offline evaluation for QAC, we focus only on standard approaches for QAC. While we do not discard the use of more complex methods, such as $[1, 3, 6, 10, 11]$ (which are left as future work), we opted for methods that are fast enough to operate in real time, a requirement of production environments. The methods used are:

Most Popular Completion (MPC) is an effective method for QAC [1] often used as a strong baseline when comparing QAC approaches $[2, 3, 5, 6, 10]$.

Co-occurrences on Queries (COQ) is a fast method based on an inverted index of past queries. Like MPC, past queries are saved into a database and used to suggest completions. When completions are required, COQ tokenizes the current query, issuing a Boolean request to retrieve all past queries that have all the keywords used in the user query. Then, the most frequent words in the result set are recommended to the user. In the example *asthma children t*, we would filter the past queries first, retaining only those containing both *asthma* and *children*, then we would order the most frequent words that start with *t*.

Co-occurrences on Titles (COT) is similar to COQ, however titles of past clicked documents are used instead of past queries.

In order to implement these methods, we take advantage of the historical clicks of *Trip Database*[3], a commercial medical search engine. We collected a

[3] https://www.tripdatabase.com

sample of 1.3 million clicks from November 2010 to February 2015, from which the vast majority (around 80%) are recent logs from January 2014 to February 2015. Each entry of these click logs has information regarding: (1) the query issued by a user, (2) the time a user clicked on a document, and (3) the title of the clicked document.

3 Experiments and Results

We trace here parallel experiments using standard online and offline evaluation procedures. Our goal is to understand the insights that each evaluation method would bring us.

During a period of 3 weeks in Sep./Oct. 2016, we collected clicks on query completions while the user was typing the query. In each week, a comparison of two different QAC approaches was done: in the first week we compared MPC and COQ (M-Q), in the second, MPC and COT (M-T), and in the third, COT and COQ (Q-T). Whenever a user clicked on a completion, the interaction was saved into a log file. This data was used for the online evaluation (Section 3.1) and posteriori offline evaluation (Section 3.3). For the same period of time, we got access to the click logs containing all the clicks made by users (as described in Section 2). This data is used for the offline evaluation (Section 3.2).

3.1 Online Evaluation

Our online evaluation was done with interleaving. We used a modification of the Team-Draft Interleaving (TDI) algorithm [9], which does not assign a team to the top common elements of both lists [4]. This particular modification of TDI was shown to further increase the sensitivity of the TDI algorithm.

In Figure 1, we show the results aggregated by experiment day for our three-week comparison between each pair of methods described in Section 2. MPC is the dominating method, performing better than COQ and COT across all the days. Note also that COQ outperforms COT.

3.2 Offline Evaluation

At the same period of the online evaluation, we collected 55,805 document clicks made from 27,040 unique queries[4]. These 55k queries were used in the offline evaluation, as described in Section 1. The evaluation procedure is the following: for each query, a prefix of length L is used to generate completions using MPC, COQ and COT. The Mean Reciprocal Rank (MRR) is then used to evaluate each method. Figure 2 shows the MRR score for each metric with L varying from 2 to 20[5].

We also allocated all 55K queries into the 3 weeks according to the period in which they were issued. This aims to evaluate the bias regarding the experiment that was in place. For example, in the first week, when MPC was compared to COQ, would it be fair to compare these two methods with another one, such as COT, that was not part of this experiment? Table 1 shows the results of

[4] An average of 2.06 documents were clicked per query. Queries without document clicks were not recorded.

[5] Note that in our online experiments, the average query length on which the users clicked as a completion is 11 characters.

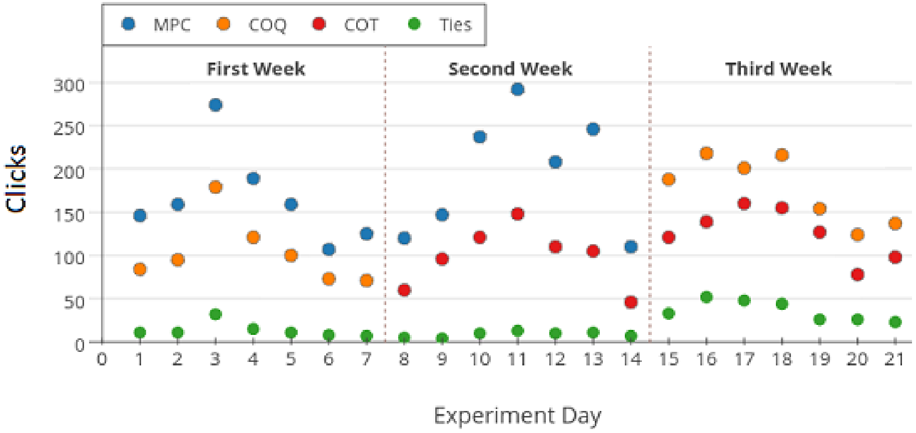

Fig. 1: Online paired interleaving evaluation made in 3 weeks: each week a different pair was compared. *Ties* refer to clicked completions on which no team was assigned.

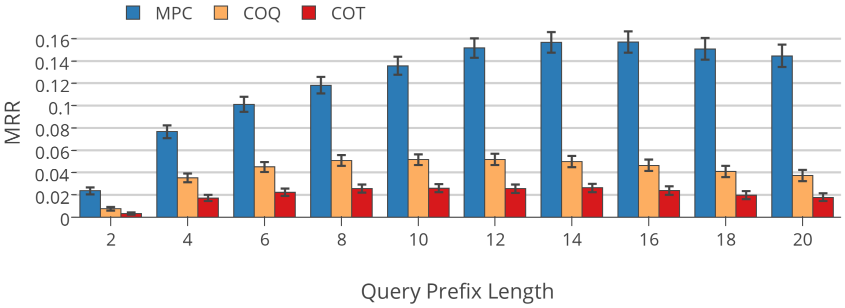

Fig. 2: Offline evaluation: mean reciprocal rank for different prefix lengths. All 3 methods are tested. Error bars show 95% confidence interval.

this experiment for different query lengths. Note that, although with varying scores for MRR, MPC consistently outperforms COQ and COT. However, the comparison between COQ and COT, which are similar methods, have a major dependence on the kind of experiment that was in place. When comparing M-Q (data from the first week MPC vs COQ) for any prefix length, we see that COQ clearly outperforms COT, however in M-T weeks we cannot say anymore that COQ outperforms COT.

3.3 Offline Evaluation Using Only Online Data

In the last part of our experiments, we explore even further the bias of the offline comparison of different methods. This time, we use only the data produced during our online experiment to perform the offline experiment. This would be

Offline evaluation using all the queries issued																
	Prefix length 4 chars.				Prefix length 10 chars.				Prefix length 16 chars.				Average over all prefixes			
	M-Q	M-T	Q-T	All	M-Q	M-T	Q-T	All	M-Q	M-T	Q-T	All	M-Q	M-T	Q-T	All
MPC	0.077	0.067	0.057	0.066	0.136	0.117	0.097	0.115	0.157	0.134	0.108	0.130	0.118	0.102	0.083	0.099
COQ	0.035	0.025	0.029	0.030	0.052	0.032	0.047	0.043	0.047	0.026	0.042	0.038	0.041	0.026	0.037	0.035
COT	0.017	0.020	0.020	0.019	0.026	0.031	0.035	0.031	0.024	0.029	0.032	0.029	0.021	0.025	0.028	0.025

Table 1: MRR scores using all the data collected during the 3 weeks of our experiments (Section 3.2).

Offline evaluation using only clicked-completion queries																
	Prefix length 4 chars.				Prefix length 10 chars.				Prefix length 16 chars.				Average over all prefixes			
	M-Q	M-T	Q-T	All	M-Q	M-T	Q-T	All	M-Q	M-T	Q-T	All	M-Q	M-T	Q-T	All
MPC	0.303	0.324	0.100	0.246	0.508	0.476	0.209	0.397	0.611	0.620	0.296	0.500	0.455	0.451	0.200	0.367
COQ	0.172	0.113	0.162	0.148	0.232	0.164	0.254	0.216	0.177	0.113	0.203	0.165	0.182	0.125	0.196	0.168
COT	0.085	0.107	0.128	0.106	0.102	0.164	0.200	0.156	0.069	0.131	0.152	0.119	0.082	0.126	0.151	0.120

Table 2: MRR scores when only using the queries that received a click as a completion during the 3 weeks of our experiments (Section 3.3).

the equivalent of using the data generated in campaigns such as CLEF Living Labs[6] and TREC-opensearch[7] to test a new algorithm after these campaigns stop running.

Table 2 shows the result of restricting the offline evaluation to only use the online data produced in the online evaluation. Here we can see how strong the bias towards the methods used in the online data is. For example, consider the week in which COQ and COT were compared (Q-T), if another method such as MPC were tested with the data generated in this week, we would probably say that this is not a good method, as for both small and average prefix length (4 and 10 characters), MPC was outperformed by COQ and COT.

4 Discussion and Conclusion

In this paper, we performed and analysed a full online evaluation comparing three query auto completion methods throughout a period of three weeks. Our online experiment collected 6,014 clicks and shows that users systematically prefer the MPC method over COQ and COT, and prefer COQ over COT.

Note that most of the commercial search engines already have a QAC method running. It means that there is always an existing untold bias towards the system in production when data is collected to create query logs for offline evaluation of new QAC methods. We evaluate this bias in two different offline experiments, using all 55k queries issued in the period of the experiment and using only the queries that were clicked on as completion during the period of the experiment. The first offline experiment, performed using a standard approach in the literature (Figure 2), produces the same results as the online evaluation produces: MPC is the best, followed by COQ, and COT last. However, when breaking

[6] living-labs.net/clef-lab/

[7] http://trec-open-search.org/

this analysis into the 3 different weeks (therefore different QAC methods in the production system), we noticed that similar methods, such as COQ and COT, are harder to tell apart. The data for the weeks in which we were comparing MPC and COT cannot be used to compare COQ as, depending on the prefix length of the queries, COT might outperform COQ, which we know should not happen. The biggest bias is found for comparisons using only the online queries in an offline manner. There we saw that for query lengths of 4 and 10, MPC is the worst method if the data comparing COQ and COT is used. This result is highly undesirable as the development of good methods such as MPC would be impacted by the bias present in the data used.

A major implication of this work is that, although we did not directly use the data created in live campaigns such as CLEF Living and TREC-opensearch, our experiments show that an extra care should be taken when using such data after the evaluation period (in an offline fashion), in order to control the bias towards the methods used in the live system. A way to mitigate this bias is by adding unbiased data, such as additional user clicks, as shown through our experiments.

Acknowledgment

This research was partially supported by the EU Project KConnect (Grant No.: 644753) and the Austrian FWF Project ADmIRE (Project No.: P25905-N23).

References

1. Ziv Bar-Yossef and Naama Kraus. Context-sensitive query auto-completion. In *Proc. of WWW*, 2011.
2. Fei Cai and Maarten de Rijke. Selectively personalizing query auto-completion. In *Proc. of SIGIR*, 2016.
3. Fei Cai, Shangsong Liang, and Maarten de Rijke. Time-sensitive personalized query auto completion. In *Proc. of CIKM*, 2014.
4. Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Trans. Inf. Syst.*, pages 1–41, 2012.
5. Giovanni Di Santo, Richard McCreadie, Craig Macdonald, and Iadh Ounis. Comparing approaches for query autocompletion. In *Proc. of SIGIR*, 2015.
6. Jyun-Yu Jiang, Yen-Yu Ke, Pao-Yu Chien, and Pu-Jen Cheng. Learning user reformulation behavior for query auto-completion. In *Proc. of SIGIR*, 2014.
7. Thorsten Joachims. Evaluating retrieval performance using clickthrough data. *Text Mining*, pages 79–96, 2003.
8. Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, pages 140–181, 2009.
9. Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How does clickthrough data reflect retrieval quality? In *Proc. of CIKM*, 2008.
10. Milad Shokouhi. Learning to personalize query auto-completion. In *Proc. of SIGIR*, 2013.
11. Milad Shokouhi and Kira Radinsky. Time-sensitive query auto-completion. In *Proc. of SIGIR*, 2012.