

Crowdsourcing Interactions

Capturing query sessions through crowdsourcing

G. Zuccon*, T. Leelanupab†, S. Whiting*,
E. Yilmaz*, J. M. Jose*, and L. Azzopardi*

*University of Glasgow, UK; *Microsoft Research, Cambridge, UK

† King Mongkut’s Institute of Technology Ladkrabang, Thailand

Abstract. The TREC evaluation paradigm, developed from the Cranfield experiments, typically considers the effectiveness of information retrieval (IR) systems when retrieving documents for an isolated query. A step forward towards a robust evaluation of interactive information retrieval systems has been achieved by the TREC Session Track, which aims to evaluate retrieval performance of systems over query sessions. Its evaluation protocol consists of *artificially* generated reformulations of initial queries extracted from other TREC tasks, and relevance judgements made by NIST assessors. This procedure is mainly due to the difficulty of accessing session logs and because interactive experiments are expensive to conduct.

In this paper we outline a protocol for acquiring user interactions with IR systems based on crowdsourcing. We show how *real* query sessions can be captured in an inexpensive manner, without resorting to commercial query logs.

1 Introduction

Traditional information retrieval research has focused on developing models and techniques for maximising the number of relevant documents retrieved for a *single query*. In practice however, this is not the case as searching for information is usually a highly interactive process. This aspect has been long ignored or marginalised within IR research. Recent developments have however shifted the focus towards investigating the interactions between users and system during the search process, as well as studying how the systems can support such interactions (see for example [3]). The ultimate goal is to develop systems that take into account all the queries a user issues as well as their interactions occurring during a search session as to assist them throughout. Within this context, an emerging research thread focuses on how those systems can be experimentally evaluated. The TREC Session track initiative [4] aims to evaluate IR systems over query sessions through controlled laboratory experiments, i.e. the “Cranfield evaluation paradigm” (“based on the abstraction of a test collection” [6]: a set of documents, a set of topics and a set of relevance judgements from NIST experts), as opposed to interactive experiments. In doing so, much of the interaction is lost and the evaluation methodology resorts to query reformulations that have been *artificially* built [4]. For example, query reformulations of the specification type are generated by selecting a query and one of its subtopics from the TREC Web diversity task dataset and extracting appropriate keywords.

Crowdsourcing, implemented in a number of web-based platforms such as Amazon Mechanical Turk (AMT)¹ and CrowdFlower², has been used as an inexpensive and often efficient way to conduct large-scale focused studies. The crowdsourcing paradigm has been already successfully used in IR for performing a number of tasks. For example, crowdsourcing has been used to gather relevance assessments for the TREC 2010 Blog Track [5]. In this paper, we outline a protocol for capturing user interactions throughout a search session using crowdsourcing. The protocol is based on the proposal made by Zuccon et al [7].

Intuitively, the protocol for capturing search sessions information through crowdsourcing operates as follows: Workers (i.e. users of the crowdsourcing tool) are asked to complete information seeking tasks within a web-based crowdsourcing platform. During the information seeking tasks, workers are assisted by an IR system encapsulated within the web-based crowdsourcing platform. While workers perform information seeking tasks, researchers can capture logs of workers interactions with the IR system. Furthermore, researchers have the possibility to acquire entry and post-search information and statistics, which would help to characterise (to some extent) the user population.

2 A protocol for Crowdsourcing Query Sessions

In [7] we outlined a protocol for conducting interactive IR experiments within crowdsourcing platforms. Information seeking tasks are performed within the following context. Crowdsourced workers are provided with a search engine embedded into the crowdsourcing platform which assists them acquiring information to satisfy their information need. Workers can issue an initial query, read document snippets, interact with the documents, and issue new queries or query reformulations. The protocol then develops around four major aspects: (1) characterise user population, (2) define information seeking tasks, (3) capture interactions, (4) acquire post-retrieval information. In this paper, we limit the discussion to the aspects influencing the acquisition of search session data, and in particular of query sessions. We thus focus on aspects (2) and (3).

2.1 Define Information Seeking Tasks

Information seeking tasks assigned to crowdsourced workers have to be clear and well defined, as no interaction is possible between workers and requesters. Workers are unlikely to perform the cognitive effort required by simulated situations and information seeking tasks as defined within the literature for laboratory based interactive IR experiments [1], i.e. by simulating search situations. This is because the workers' main goal is to complete tasks as efficiently and rapidly as possible. We suggest that in crowdsourced search environments, the topic of the search session has to be explicitly stated to workers, together with a number of specific informational questions they are expected to answer. In preliminary experiments conducted to test the validity of the protocol [7], we employed topics and questions selected from the TREC Question Answering (QA) Track for the years 2005, 2006, 2007 [2]. Consider for example topic 279 extracted from

¹ <http://www.mturk.com/>

² <http://crowdflower.com/>

the topic set of the TREC 2007 QA Track: “Australian wines”. With respect to this topic, workers are asked to use the provided search engine to help them answer the following questions: “What winery produces Yellowtail?” (279.4), “Where does Australia rank in exports of wine?” (279.3), and “Name some of Australia’s female winemakers” (279.5).

In [7] we argue that posing questions about a specific topic sufficiently initiates the workers’ information needs, thus avoiding the need to create simulated tasks. The scenario in which the information seeking task is performed is kept straightforward: workers have to answer a number of questions; to do so they can search for information using the provided IR system.

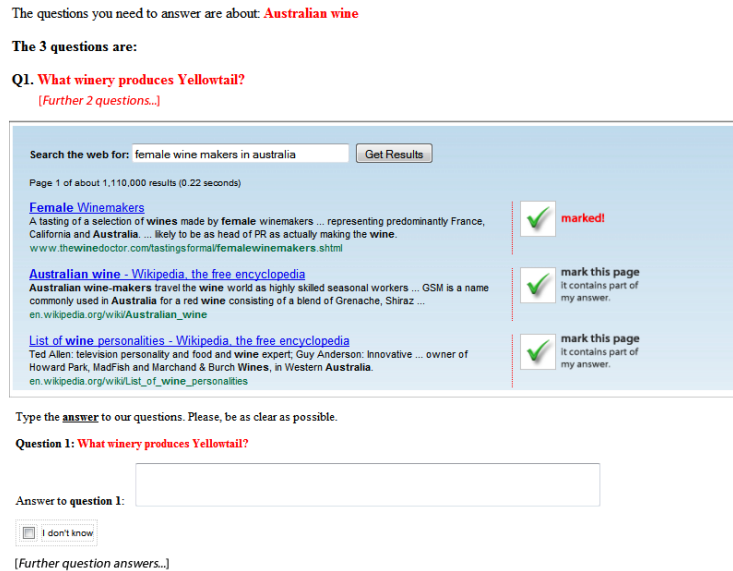


Fig. 1: The interactive search interface encapsulated within an AMT HIT.

2.2 Capture Interactions

Once topics and questions are assigned, workers can search for answers using the provided IR system. The IR system is able to record the interactions between the workers and the system (e.g. issued queries, clicked results, time spent in reading/searching, etc). Crowdsourcing platforms, such as AMT, do not provide native tools for capturing these kind of user interactions. However, several solutions can be devised so as to direct workers towards a tool that is controlled by experimenters, and thus records workers’ interactions. For example, a proxy server could be used to achieve this goal. An alternative solution can be developed as follows: workers are shown the interface of the IR system within a self-contained iFrame positioned in the page of the HIT. Through iFrames, interactions could be recorded, making them available for further analysis. We adopted the latter solution when developing our preliminary experiments. The platform of the experimental system is shown in the next section, together with an example of the interactions we were able to capture during our preliminary experiments.

3 Crowdsourcing search sessions: an example

We embedded a search system within the crowdsourcing platform offered by AMT, using an iFrame within a standard HIT. Our IR system was developed as a web-based front-end of the Microsoft Bing API³ for web results. Each time a user began a search task our system was provided with AMT HIT details such as the work assignment ID and the corresponding question topic. Queries, result clicks and explicit feedback via an optional “Mark as Relevant” button were logged alongside the HIT information. Following completion of the batch of HITs for each experiment we then merged the provided search logs with the AMT logs to yield a rich source of individual worker data for analysis⁴. AMT data provided statistics such as the search task duration, question answers, unique worker IDs and qualification scores that can be used to begin explaining behaviours observed through the related query session logs. A screenshot of the search interface encapsulated within an AMT HIT is given in Fig. 1.

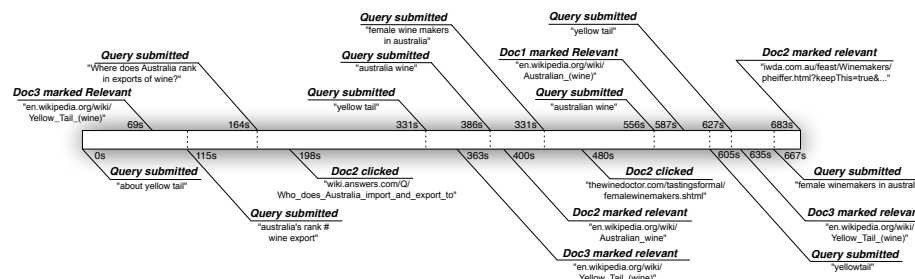


Fig. 2: Search session of a crowdsourced worker for the topic “Australian wines”.

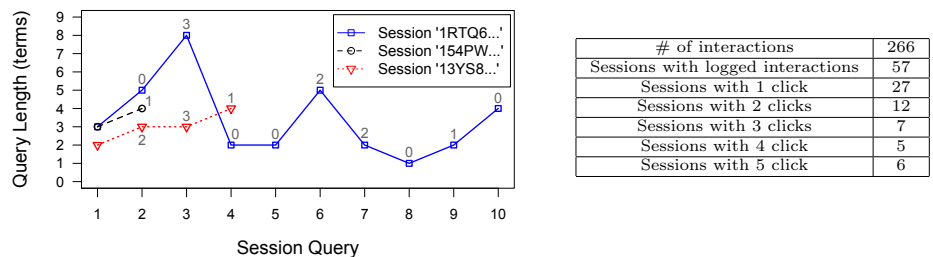
In Fig. 2 we report a session we observed during our experiments. The acquired interaction refers to the example topic of section 2.1 (“Australian wines”). During this search session, that lasted about 11 minutes, the worker issued 10 queries aiming to solve his information need, i.e. answer the three questions we asked about Australian wines. The worker also observed the snippets of the retrieved documents, clicked to access some of the documents and marked some of these as useful (i.e. relevant) for solving the information task. Within the session, the worker issued all the three types of queries recognised by the TREC Session Track: specification (“australia wine” → “female wine makers in australia”), drifting (“about yellow tail” → “australia’s rank # wine export”), and generalisation (“yellow tail” → “australia wine”). In Fig. 3a we report the evolution in terms of query length of three of the sessions recorded for topic 279.

The statistics regarding query sessions collected during our experiments are as follows. We collected in total 119 queries for 24 HITs (with 58 assignments completed by workers). Users were encouraged to engage with the retrieval system by promising the award of a bonus payment if they did so: while, their HIT was refused if questions were answered without interacting with the search engine. A qualification test based on aptitude tests (as proposed in [7]) was used to characterise the user population. Excluding workers not issuing any queries (1

³ <http://www.bing.com/developers>

⁴ Logs and search interface’s screenshots are available at <http://www.dcs.gla.ac.uk/~guido/sir2011>.

session), workers issued on average 2.26 queries per session (the average length of search sessions is about 5.45 minutes). The queries are on average 3.78 terms long, with a maximum of 10 terms (including stopwords). Finally, in Fig. 3b we report statistics of the interactions we captured during our experiments.



(a) Evolution in terms of query length of the three sessions recorded for topic 279 of the TREC 2007 QA Track. Labels indicate the number of shared terms between new and previous query. (b) Statistics regarding the interactions captured during our experiments.

Fig. 3

4 Directions of Development

In this paper we have shown how crowdsourcing can be used to capture search sessions and in particular query sessions. We have suggested that crowdsourcing allows researchers to obtain search session data, and in particular query sessions, which are more akin to represent real users' queries than those generated within the TREC 2010 Session Track. We also have outlined a protocol that can be used within the TREC Session Track for crowdsourcing real query sessions to be used for evaluating interactive IR systems. Future work will be directed towards the consolidation and evaluation of the introduced crowdsourcing protocol, in particular by comparing the acquired information against that obtained through laboratory based experiments and session logs obtained from commercial search engines. Furthermore, we intend to explore the possibility of fully relying on crowdsourcing for the evaluation of interactive IR systems.

References

1. P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 2003.
2. H. T. Dang, et al. Overview of the TREC 2007 Question Answering Track. *TREC'07*, 2007.
3. N. Fuhr. A probability ranking principle for interactive information retrieval. *JIR*, 12(3):251–265, June 2008.
4. E. Kanoulas, et al. Session track at trec 2010. In *Proc. of SimInt 2010*, 2010.
5. R. McCreadie, et al. Crowdsourcing blog track top news judgments at trec. In *Proc. of CSDM 2011*, 2011.
6. E. M. Voorhees. Trec: Improving information access through evaluation. *Bulletin of the American Society for Information Science and Technology*, 32(1):16–21, 2005.
7. G. Zuccon, et al. Crowdsourcing Interactions: A proposal for capturing user interactions through crowdsourcing. In *Proc. of CSDM 2011*, pages 35–39, 2011.