Report on the ECIR 2011 Workshop on Information Retrieval Over Query Sessions

Ben Carterette
University of Delaware
carteret@cis.udel.edu

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

Evangelos Kanoulas
University of Sheffield
e.kanoulas@sheffield.ac.uk

Mark Sanderson
RMIT University
mark.sanderson@rmit.edu.au

Abstract

Research in Information Retrieval has traditionally focused on serving the best results for a single query. Real users, however, often begin an interaction with a search engine with a sufficiently under-specified information need that they will need to reformulate before they find either the one thing or every thing they are looking for. We define a *session* as the sequence of queries and interactions that a user performs in service of an information need. The first workshop on Information Retrieval Over Query Sessions was held at ECIR 2011 in Dublin, Ireland, with the purpose of investigating questions of measuring, analyzing, and optimizing IR system behavior over a session of reformulations.

1 Introduction

The idea of supporting information retrieval (IR) over user sessions has not received as much attention in IR as one might expect, despite the interactive and dynamic nature of real life searching. However, the growing interest in Interactive IR (IIR), coupled with recent studies of sessions (and trails) within transaction logs and the organisation of the TREC Session Track, has seen the topic gain considerable momentum in recent years. It is commonly accepted that modeling and evaluating IR as single and independent queries does not faithfully reflect the dynamic and iterative nature of real life searching. Most users will issue multiple queries that form part of a search session involving an iteration of user-system interactions. This might be the result of users clarifying or refining their information needs (e.g. broadening or narrowing a search), in response to the search results (e.g. too few or too many hits). Early studies on web search query logs showed that half of all Web users reformulated their initial query: 52% of the users in 1997 Excite data set, 45% of the users in the 2001 Excite data set. A search engine may be able to better serve a user by considering the user interaction with the search results (query reformulations, clicks, dwell times), employing techniques of query disambiguation, query recommendation, search personalization, focused diversification, or other currently unanticipated techniques.

The first workshop on Information Retrieval Over Query Sessions was held at ECIR 2011 in Dublin, Ireland, with the purpose of providing a forum to discuss ideas on measuring, analyzing, and optimizing IR system behavior over a session of reformulations. The workshop was structured along two main themes:

1. Algorithms that use information provided throughout query sessions for a number of tasks, ranging from ranking to query suggestion and construction of user profiles,
2. Query session detection and construction of query session collections.

The workshop grew out of the first TREC Sessions track, which was completed in October 2010, and subsequent discussions on the right way to model the problem for future experimental evaluation as well as the right way to evaluate it with test collections.

The workshop took place 18th April 2011 in the Golden Barley Room at the Guinness Storehouse. It was one of two full-day workshops collocated with the conference. Attendance was strong with more than 20 international participants.

2 Workshop programme and summary

The programme consisted of a keynote talk, presentations of nine research papers, and a panel. Below we summarize some of the main points from the presentations and discussion.

2.1 Keynote

Rosie Jones of Akamai opened the workshop with a keynote talk entitled "Behavioural Targeting and Search Sessions" on the applications of analysis of sessions in web query logs. Dr. Jones gave an overview of the relationship between display advertising and search advertising, and discussed how the analysis of sessions within transactions logs could be used to assist with search marketing. She presented a model of sessions in which a user's interaction with a system comprises multiple *missions*, each of which consist of discrete *goals* for which the user formulates individual queries. An example of a mission is a user looking for places to hike near San Francisco. Goals include finding lists of state parks, finding lists of hiking trails, and finding information about particular state parks or hiking trails. In general, queries and goals may be interleaved amongst each other and between missions, presenting a challenge for analyzing sessions using only query logs.

Dr. Jones then discussed the application of this model of sessions to shopping and similar tasks. For example, the distribution of queries correlated with the first appearance of the query "mortgage" changes over time: initial queries relate to mortgage calculators, financing, and lenders; a week later the distribution shifts to queries about homes for sale, builders, appraisers, and realtors. Within a month the distribution shifts to queries about insurance, legal matters, and furniture. Three months later, queries largely relate to furnishing and decorating, in particular for large retailers like Kohl's, Sears, and Pottery Barn. Finally, a year out, we see queries about outdoor furniture, pools, lawns, and other such things. These observations present opportunities for detecting and predicting *what* to advertise to particular users and demographics and *when* to present those ads.

Dr. Jones discussed methods for segmenting sessions into missions and goals, and methods for predicting whether missions and goals were successful. Doing this well is probably necessary to be able to successfully leverage log data for search advertising. But it is a very hard problem in general.

2.2 Paper sessions

Nine papers were mostly presented in two sessions, one in the morning and one in the afternoon. Below we summarize some of the key findings as they relate to the two main themes of the workshop.

2.2.1 Using interaction information to improve search engines

There were five papers on the use of interaction information to improve some aspect of a search engine, whether it be retrieval effectiveness, effectiveness for personalization, or alternate query suggestions. These papers covered a variety of domains and tasks, including ad hoc retrieval, shopping, and product rating.

M-Dyaa Albakour presented *The Use of Domain Modelling to Improve Performance Over a Query Session*, co-authored with Deirdre Lungley and Udo Kruschwitz [8]. This work was about improving retrieval effectiveness for the last query in a session given previous queries and information mined from ranked results, specifically semantic concepts as determined by Formal Content Analysis (FCA) of ranked results for previous queries. Using data from the TREC 2010 Sessions track, they demonstrated a gain in effectiveness when expanding the last query using these concepts over using no prior information at all.

The next paper was given by Corrado Boscarino titled *Implicit relevance feedback from a multi-step search process: a use of query-log* (with Arjen de Vries, Vera Hollink, and Jacco van Ossenbruggen) [1]. The paper uses implicit feedback in the form of user clicks to expand a query given previous queries and interactions. Though a click is not necessarily a relevance judgment, there is some evidence that such implicit feedback may improve results in an image purchasing task.

The third presentation of the morning session was for the paper *New user profile learning for extremely sparse data sets* co-authored by Tomasz Hoffmann, Tadeusz Janasiewicz, and Adrzej Szwabe [5]. This work presents a new method for estimate user profiles given sparse interaction data. Using their method to predict movie ratings, there was a substantial improvement over other methods from the literature.

The final presentation of the morning session was for Hao Wu and Hui Fang's paper *An Exploration of Query Term Deletion*, presented on their behalf by Ben Carterette [9]. The authors argue that systems may be able to improve effectiveness for a user's query by deleting some terms. Using the TREC 2010 Sessions data (specifically pairs of queries for which the second used a subset of terms in the first), they show that a few automatic heuristics for deleting terms can improve effectiveness over the manual deletion done for the track.

In the afternoon session, Makoto Kato presented a paper titled *Query Session Data vs. Clickthrough Data as Query Suggestion Resources* co-authored with Tetsuya Sakai and Katsumi Tanaka [6]. They compared the use of query logs and click logs to generate query suggestions. Using real data from Microsoft Bing, they showed that query log analysis tends to lead to much better suggestions than click data.

2.2.2 Detecting sessions and building session test collections

There were four papers on the general themes of detecting sessions in query logs and building test collections for studying session retrieval.

In the afternoon session, Johannes Leveling presented work with Gareth Jones entitled *Same Query – Different Results? A Study of Repeat Queries in Search Sessions* [7]. An

analysis of a large query log suggests that repeat queries actually make up the plurality of reformulations, and that users are actually frequently expecting to see different results upon resubmission of a query. This has clear implications for building test collections to study session effectiveness, in that a full model of the problem should include such common cases. The Session track in its current form cannot model this.

For the first presentation of the afternoon session, Matthias Hagen presented work with Bruno Stein and Tino Rüb entitled *Query Session Detection as a Cascade* [4]. This work presents a method based on cascading features of increasing complexity for computationally-efficient on-line detection of query sessions. They show that they can accurately segment a log into sessions while also saving time over state-of-the-art methods.

Next, Johannes Leveling presented a second paper with Debasis Ganguly and Gareth Jones on *Automatic Generation of Query Sessions using Text Segmentation* [3]. This work presents a model for generating query reformulations given documents retrieved for a previous query. It generates both "specialization" and "generalization" reformulations by adding and deleting terms respectively. This work suggests that a session test collection could be built by simulating reformulations.

Finally, immediately before the panel discussion, Guido Zuccon presented work with Teerapong Leelanupub, Stewart Whiting, Emine Yilmaz, Joemon Jose, and Leif Azzopardi on using crowdsourcing to capture sessions of interactions [10]. The presentation initiated an interesting discussion on how to generate query session datasets for the purpose of evaluation and in a sense introduced the topic of the panel discussion.

2.3 Panel discussion

The workshop concluded with a panel discussion entitled "Information Retrieval Evaluation over Query Sessions" featuring Kalervo Järvelin (University of Tampere, Finland), Stephen Robertson (Microsoft Research Cambridge, UK), Tetsuya Sakai (Microsoft Research Asia, China) and Mark Sanderson (RMIT University, Australia).

To set the stage for the discussion, Ben Carterette (University of Delaware) presented a brief overview of outcomes and conclusions from the TREC 2010 Sessions track and some directions for the 2011 track. The 2011 Sessions track data consists of "crowdsourced" sessions similar to (but not exactly the same as) those described by Zuccon in the preceding talk. This means participants have access to "real" user reformulations as well as user clicks and dwell times.

Mark Sanderson, drawing on the experience of building the 2010 TREC Session track test collection, discussed the difficulties with creating a session-based collection, in particular determining how to reliably capture sessions and how to select and use an appropriate user model. One example of the challenge of determining the user model was given by Sanderson when he described the discussion amongst session track organizers on how to deal with duplicate documents retrieved across the different queries of a session. Would users prefer to see the same documents retrieved across a session or would that find that annoying? Sanderson's point is that at present no one knows the answers to such questions.

Kalervo Järvelin suggested using medical case search as a starting point for sessions and discussed his SimInt workshop paper comparing real-life strategies on TREC queries. The generalization/drifting/specification queries in the Session Track were felt to be ambiguous compared to what users do in reality.

Stephen Robertson brought up the point that the Sessions track does not model many of

the scenarios Rosie Jones used as examples in her talk, such as specializations of query reformulation and relevance judgements do not easily cascade in such cases. Modeling interactive IR is quite complex, but the user model underlying a test collection must be kept simple in order to limit the degrees of freedom in evaluation.

Tetsuya Sakai, having come from giving a keynote at the other full-day workshop on diversity, drew connections between sessions and the diversity ranking task. He discussed the U-measure: probability distribution over sequences of atomic actions (type-type-read-read-click-end) in which the cost is the sequence of atomic actions and the benefit of the accumulated nuggets (i.e. not just about document relevance). In his view accumulating data to answer questions gathered across sessions would be an appropriate measure of success.

3 Conclusions and future directions

The first workshop on Information Retrieval Over Query Sessions seems to have succeeded in its goal of providing a forum for discussion about analyzing, optimizing, and evaluation retrieval systems over a user session. While it is clear that the problem is hard, it also seems that there are many opportunities for solid research results. The TREC Session track in its current form is modeling one relatively simple aspect of the problem as a whole, which may suggest the need for future evaluation workshops and test collections that can consider different aspects.

References

[1] Corrado Boscarino, Arjen P. de Vries, Vera Hollink, and Jacco van Ossenbruggen. Implicit relevance feedback from a multi-step search process: a use of query-logs. In Carterette et al. [2].

[2] Ben Carterette, Evangelos Kanoulas, Paul D. Clough, and Mark Sanderson, editors. *Proceedings of the ECIR 2011 Workshop on Information Retrieval Over Query Sessions*, Available at `http://ir.cis.udel.edu/ECIR11Sessions`.

[3] Debasis Ganguly, Johannes Leveling, and Gareth Jones. Automatic generation of query sessions using text segmentation. In Carterette et al. [2].

[4] Matthias Hagen, Benno Stein, and Tino Rüb. Query session detection as a cascade. In Carterette et al. [2].

[5] Tomasz Hoffman, Tadeusz Janasiewicz, and Adrzej Szwabe. New user profile learning for extremely sparse data sets. In Carterette et al. [2].

[6] Makoto P. Kato, Tetsuya Sakai, and Katsumi Tanaka. Query session data vs clickthrough data as query suggestion resources. In Carterette et al. [2].

[7] Johannes Leveling and Gareth Jones. Same query – different results? a study of repeat queries in search sessions. In Carterette et al. [2].

[8] Deirdre Lungley, M-Dyaa Albakour, and Udo Krushwitz. The use of domain modeling to improve the performance over a query session. In Carterette et al. [2].

[9] Hao Wu and Hui Fang. An exploration of query term deletion. In Carterette et al. [2].

[10] Guido Zuccon, Teerapong Leelanupab, Stewart Whiting, Emine Yilmaz, Joemon Jose, and Leif Azzopardi. Crowdsourcing interactions: Capturing query sessions through crowdsourcing. In Carterette et al. [2].