

Session Track at TREC 2010

Evangelos Kanoulas*

Paul Clough*

Ben Carterette†

Mark Sanderson*

*Department of Information Studies
University of Sheffield
Sheffield, UK

†Department of Computer & Information Sciences
University of Delaware
Newark, DE, USA

1. INTRODUCTION

Research in Information Retrieval has traditionally focused on serving the best results for a single query, e.g. the most relevant results, a single most relevant result, or a facet-spanning set of results. In practice, no matter the task, users often enter a sufficiently ill-specified query that one or more reformulations are needed in order to locate a sufficient number of what they seek. 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 dataset [9]. A search engine may be able to better serve a user not by ranking the most relevant results to each query in the sequence, but by ranking results that help “point the way” to what the user is really looking for, or by complementing results from previous queries in the sequence with new results, or in other currently-unanticipated ways.

The standard evaluation paradigm of controlled laboratory experiments is unable to assess the effectiveness of retrieval systems to an actual user experience of querying with reformulations. On the other hand, interactive evaluation is both noisy due to the high degrees of freedom of user interactions, and expensive due to its low reusability and need for many test subjects. In this work we propose an initial experiment that can be used to evaluate the simplest form of user contribution to the retrieval process, a single query reformulation. This experiment is the basis of the TREC 2010 Session track.

2. EVALUATION TASKS

We call a sequence of reformulations in service of satisfying an information need a session, and the goals of our evaluation are: (**G1**) to test whether systems can improve their performance for a given query by using information about a previous query, and (**G2**) to evaluate system performance over an entire query session instead of a single query. We limit the focus of the track to sessions of two queries, and further limit the focus to particular types of sessions (defined

below). This is partly for pragmatic reasons regarding the difficulty of obtaining session data, and partly for reasons of experimental design and analysis: allowing longer sessions introduces many more degrees of freedom, requiring more data from which to base conclusions.

A set of 150 query pairs (original query, query reformulation) is provided to TREC participants. For each such pair the participants are asked to submit three ranked lists of documents for three experimental conditions, (a) one over the original query (**RL1**), (b) one over the query reformulation, ignoring the original query (**RL2**), and (c) one over the query reformulation taking into consideration the original query and its search results (**RL3**). By using the ranked lists (RL2) and (RL3) we evaluate the ability of systems to utilize prior history (G1). By using the returned ranked lists (RL1) and (RL3) we evaluate the quality of ranking function over the entire session (G2). Note that in contrast to the Relevance Feedback track this is a one-phase evaluation with no feedback provided by assessors.

3. QUERY REFORMULATIONS

There is a large volume of research regarding query reformulations which follows two lines of work: a descriptive line that analyzes query logs and identifies a taxonomy of query reformulations based on certain user actions over the original query (e.g. [6, 1]) and a predictive line that trains different models over query logs to predict good query reformulations (e.g. [4, 3, 8, 5]). Analyses of query logs showed a number of different types of query reformulations with three of them being consistent across different studies (e.g. [4, 6]):

Specifications: the user enters a query, realizes the results are too broad or that they wanted a more detailed level of information, and reformulates a more specific query.

Drifting/Parallel Reformulation: the user entered a query, then reformulated to another query with the same level of specification but moved to a different aspect or facet of their information need.

Generalizations: the user enters a query, realizes that the results are too narrow or that they wanted a wider range of information, and reformulated a more general query.

In the absence of query logs, Dang and Croft [2] simulated query reformulations by using anchor text, which is readily available. In this work we use a different approach. To construct the query pairs (original query, query reformulation) we start with the TREC 2009 Web Track diversity topics. This collection consists of topics that have a “main theme” and a series of “aspects” or “sub-topics”. The Web Track queries were sampled from the query log of a commer-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$10.00.

cial search engine and the sub-topics were constructed by a clustering algorithm [7] run over these queries aggregating query reformulations occurring in the same session. We used the aspect and main theme of these collection topics in a variety of combinations to provide a simulation of an initial and second query. An example of part of a 2009 Web track query is shown below.

```
<topic number="4" type="faceted">
  <query>toilet</query>
  <description> Find information on buying, installing,
    and repairing toilets.
  </description>
  <subtopic number="1" type="inf">
    What different kinds of toilets exist, and how do
    they differ?
  </subtopic>
  ...
  <subtopic number="3" type="inf">
    Where can I buy parts for American Standard toilets?
  </subtopic>
  ...
  <subtopic number="6" type="inf">
    I'm looking for a Kohler wall-hung toilet. Where can
    I buy one?
  </subtopic>
</topic>
```

To construct **specification** reformulations we used the Web Track `<query>` element as the original query, selected a subtopic and considered it as the actual information need. We then manually extracted keywords from the sub-topic and used them as the reformulation. For instance, in the example above we used the query “toilet” as the first query, selected the information need (“I’m looking for a Kohler wall-hung toilet. Where can I buy one?”), extracted the keyword “Kohler wall-hung” and considered that as a reformulation. This query pair simulates a user that is actually looking for a Kohler wall-hung toilet, but poses a more general query first, possibly because they don’t “know” what they need.

```
<topic number="1" reformtype="specification">
  <query>toilet</query>
  <reformulation>Kohler wall-hung toilet</reformulation>
  <description>I'm looking for a Kohler wall-hung toilet.
    Where can I buy one?
  </description>
</topic>
```

To construct **drifting** reformulations we selected two subtopics, used the corresponding `<subtopic>` elements as the description of two separate information needs, extracted keywords out of the subtopic, and used these keywords respectively as the query and query reformulation. For instance, in the example above we selected subtopics 3 and 6 as the two information needs. Then we extracted the keywords “parts American Standard” and “Kohler wall-hung toilet” and used them as the original query and the query reformulation. This pair simulates a user that first wants to buy toilet parts from American Standard and then decides that they also want to purchase Kohler wall-hungs while browsing the results.

```
<topic number="2" reformtype="drifting">
  <query>parts American Standard</query>
  <description>Where can I buy parts for
    American Standard toilets?</description>
  <reformulation>Kohler wall-hung toilet</reformulation>
  <rdescription>I'm looking for a Kohler
    wall-hung toilet. Where can I buy one?
  </rdescription>
</topic>
```

Finally, to construct **generalization** reformulations we followed one of two methods. In the first method we selected one of the subtopics and we extracted as many keywords as possible to construct an over-specified query, e.g. from subtopic 1 of the example topic we may extract the keywords “different kinds of toilets”, which seems to be a lexical over-specification. We then used a subset of these keywords to generalize the original query (e.g. “toilet”). This is meant to simulate a user that first wants to find what types of toilets exist, but lexically over-specifies the need; the retrieved results are expected to be poor and therefore the user needs to reformulate.

```
<topic number="3" reformtype="generalization">
  <query>different kinds of toilets</query>
  <reformulation>toilets</reformulation>
  <description> What different kinds of toilets
    exist, and how do they differ?</description>
</topic>
```

For the second method we selected one of the subtopics or the query description from the Web Track topics as the information need, extracted keywords from a different subtopic that seemed related but essentially it was a mis-specification of something very narrow, and extracted keywords from the subtopic used as information need.

```
<topic number="4" reformtype="generalization">
  <query>American Standard toilet</query>
  <reformulation>toilet</reformulation>
  <description>Find information on buying,
    installing, and repairing toilets.</description>
</topic>
```

4. CONCLUSIONS

Simulating a user is a difficult task. A test collection and accompanying evaluation measures already provide a rudimentary simulation of such users. We have chosen to extend this by considering one more aspect of typical searchers, their reformulation of a query.

5. REFERENCES

- [1] P. Bruza and S. Dennis. Query reformulation on the internet: Empirical data and the hyperindex search engine. In *Proceedings of RIAO*, pages 488–500, 1997.
- [2] V. Dang and B. W. Croft. Query reformulation using anchor text. In *Proceedings of WSDM*, pages 41–50, 2010.
- [3] J. Huang and E. N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of CIKM*, pages 77–86, 2009.
- [4] B. J. Jansen, D. L. Booth, and A. Spink. Patterns of query reformulation during web searching. *Journal of American Society for Information Science and Technology*, 60(7):1358–1371, 2009.
- [5] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *Proceedings of WWW*, 2006.
- [6] T. Lau and E. Horvitz. Patterns of search: analyzing and modeling web query refinement. In *UM '99: Proceedings of the seventh international conference on User modeling*, pages 119–128, 1999.
- [7] F. Radlinski, M. Szummer, and N. Craswell. Inferring query intent from reformulations and clicks. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 1171–1172, New York, NY, USA, 2010. ACM.
- [8] X. Wang and C. Zhai. Mining term association patterns from search logs for effective query reformulation. In *Proceedings of CIKM*, pages 479–488, 2008.
- [9] D. Wolfram, A. Spink, B. J. Jansen, and T. Saracevic. Vox populi: The public searching of the web. *JASIST*, 52(12):1073–1074, 2001.