

Information Retrieval Journal manuscript No.
(will be inserted by the editor)

A Term-Based Methodology for Query Reformulation Understanding

Marc Sloan · Hui Yang · Jun Wang

Received: 15 August 2014 / Accepted: 24 February 2015 / Published Online: 6 March 2015

Abstract Key to any research involving session search is the understanding of how a user's queries evolve throughout the session. When a user creates a query reformulation, he or she is consciously retaining terms from their original query, removing others and adding new terms. By measuring the similarity between queries we can make inferences on the user's information need and how successful their new query is likely to be. By identifying the origins of added terms we can infer the user's motivations and gain an understanding of their interactions.

In this paper we present a novel term-based methodology for understanding and interpreting query reformulation actions. We use TREC Session Track data to demonstrate how our technique is able to learn from query logs and we make use of click data to test user interaction behavior when reformulating queries. We identify and evaluate a range of term-based query reformulation strategies and show that our methods provide valuable insight into understanding query reformulation in session search.

Keywords Term Model · Click Model · Query Reformulation

M. Sloan
University College London, UK
E-mail: M.Sloan@cs.ucl.ac.uk

H. Yang
Georgetown University, USA
E-mail: huiyang@cs.georgetown.edu

J. Wang
University College London, UK
E-mail: J.Wang@cs.ucl.ac.uk

1 Introduction

Session search in Information Retrieval (IR) occurs when a user issues multiple queries consecutively to a search engine in the pursuit of satisfying one or more information needs. A session is typically defined as a period of continuous interaction with a search engine and can be demarcated in a number of ways, a common one being 30 minutes of inactivity (White and Drucker, 2007). Sessions containing more than one query make up a significant proportion of search activity, with one study finding 32% of sessions containing 3 or more queries (Jansen et al, 2005). Understanding the underlying interactions in session search can lead to improved search interfaces, better search rankings and user satisfaction.

Sessions are driven by query reformulations, the user controlled act of modifying an existing query in order to pursue new search results. Query reformulations are usually closely related to the user's previous query and reflect the shifting cognition of the user throughout the session search. For instance, a user may have an unclear information need at the start of a session which becomes more refined as snippets are read and documents are clicked. Such queries can be ambiguous when the user is unsure how to explicitly define his or her information need (Song et al, 2009) or explorative when the user is actively seeking a broad range of information on a subject (Marchionini, 2006). In both cases, the information need can change throughout the session, whether through specialization, generalization and so on, which leads to variations in the queries used to describe it.

We observe that sessions are typified by queries consisting of core terms related to the underlying information need and additional terms that reflect the user's cognitive changes (Kinley et al, 2012). Over the course of the session, the core terms may change as well. At any point in a session, we define three possible *term actions* available to a user:

Term Retention - Keeping terms from one query to the next, the core terms for the current information need.

Term Removal - Removing a term from a query.

Term Addition - Adding a new term not present in the preceding query to the query reformulation.

To illustrate a particular instance of query reformulation within session search and the described term actions, Table 1 contains the queries in a typical search session found in the 2013 Session Track dataset (Kanoulas et al, 2013). This session represents an explorative information need regarding public and political opinion on US gun control laws. The terms 'gun control' are **retained** through the first four queries, with the user **adding** and **removing** terms 'opinions', 'US government' and 'current affairs' in order to learn more about the topic. The focus shifts in query 5 with 'gun control' changing to 'gun violence', indicating a change in information need, which is expanded upon in the final query which is more specialized.

Table 1 Queries in session 40 of the TREC 2013 Session Track.

Impression Position	Query
1	gun control opinions
2	gun control us government
3	gun control current affairs
4	gun control current affairs
5	gun violence us
6	law center to prevent gun violence

Without knowing the underlying information need driving the queries, the example demonstrates that it is possible to infer persistent subtopics and the terms that are likely to be retained or removed from query to query (in this case ‘gun control’ and ‘gun violence’). A certain degree of overlap is typical between queries but how much? What factors influence whether a term is likely to be kept or removed in the next query? Can we determine a source for the new terms that are introduced into a query? Measuring the similarity between queries and other sources of text can help us resolve some of these questions and allow us to build descriptive and evaluative models of user behavior during a session search.

For instance, we observe in the example session that the snippets of all the results for the first query contain the terms ‘gun control’, and out of all ranked documents only the clicked document (ClueWeb ID *clueweb12-0100wb-86-17546*) contains the terms ‘US government’ (in the phrase “*US Government Info Guide*”), which were then used in the next query. One inference that could be drawn here is that the user observed the terms ‘US government’ in the clicked document which influenced their reformulation decision making process.

In this paper we seek to gain an understanding of the query reformulation process and resolve the following research questions:

1. What is the relationship between terms found in adjacent queries in search sessions. How often are terms from a query retained or removed in a query reformulation?
2. Where are query reformulation terms not present in the original query sourced from and can we model term addition?
3. Can user-behavior scenarios defined on terms that are retained, removed or added inform us of the quality of query reformulations?

We resolve these questions by introducing a novel methodology for interpreting query reformulations using terms. We use our technique to explore term retention and removal by analyzing adjacent and non-adjacent queries in sessions. With term addition, our observations indicate that a significant number of added terms in a reformulation can be sourced from the terms that the user was exposed to in the previous impression. An impression consists of a query, its snippets and its documents, all of which contain terms that the user may have encountered during session search. By also incorporating click information, we can define and evaluate three sources for such terms, *clicked* and *non-clicked snippets* and *clicked documents*.