The next stage in our analysis involved measuring the value of the three term sources in determining whether query terms were retained, removed or added, leading to eight possible scenarios of user behavior which we interpret based on our results. To evaluate the effectiveness of scenario-based term prediction, and also the user's observed query reformulations, we determine whether the term actions ultimately lead to increased user satisfaction or improved search rankings, which we measure using implicit click information and a number of IR metrics.

Our analysis was conducted on the TREC Session Track data from 2011 to 2014 (Kanoulas et al, 2011, 2012, 2013), a set of standardized query logs comprising queries grouped by sessions across a number of predefined topics, the ranked documents, their snippets and clickthroughs (including order and dwell time) and relevance judgments. The documents belong to the ClueWeb09[1] and ClueWeb12[2] corpora. This dataset was chosen as it is widely available, well regarded in the IR community and whilst small when compared to commercial query logs, is rich with potential sources for term discovery (snippets and documents), interaction data (clicks and dwell time) and relevance judgments (for evaluation).

The remainder of the paper is organized as follows. Section 2 presents the related work and Section 3 outlines the dataset used, experimental setup and the key definitions and similarity measures used in our methodology. In Sections 4 and 5 we use our term-based technique to understand the three term actions *retention*, *removal* and *addition*, investigate user click behavior and define the three term sources. In Section 6 we expand the term sources into user interaction based term scenarios and evaluate reformulation strategies. We conclude the paper and discuss our findings in Sections 7 and 8.

2 Literature Review

Session Log Analysis Ours is not the first query log analysis of query reformulation behavior. Jansen et al (2009) defined different query reformulation states and the transition patterns that occur during a session and evaluated over a large query log. Their research idea is similar to our scenario approach although in their study the states operate on a query level by looking at the degree of overlap between queries, rather than our term based approach, but some of our findings are similar. Liu et al (2010) explored a similar state-based analysis but this time on a user study that allowed them to determine different types of behavior based on the type of task being performed by the user. Kinley et al (2012) also performed a user study with the intention of observing different query modifying behavior (such as replacing, adding terms etc.) and linking it to a user's 'cognitive style' of query reformulation. A similar work to ours is Huang's (Huang and Efthimiadis, 2009) classification of different types of reformulation behavior that utilizes clicks from query logs and uses

[1] http://www.lemurproject.org/clueweb09/index.php
[2] http://www.lemurproject.org/clueweb12.php/

term differences as well. Nonetheless, ours is the first such study using a purely term-based approach that also incorporates clicks in a user interaction model.

Click and User Modeling A key component of this work is our click based methodology and our rank and impression position experiments. This is similar to work in click modeling, an established area of IR research that typically uses search logs, eye-tracking and user studies to understand how users navigate search pages. For instance, our $s_n(\text{LC})$ definition and experiments in Section 5.1 are based on the examination hypothesis model (Joachims et al, 2005; Craswell et al, 2008). In other research, eye-tracking has been used on participants with predefined search tasks, with the researchers being able to predict which task was being performed based on eye tracking patterns (Cole et al, 2011, 2010), which was further developed into being able to factor in the stage of the user's task (Liu and Belkin, 2010). Another recent eye tracking study (Liu et al, 2014) found that when browsing search results users will glance at snippets but not fully read them, returning to them at a later point if at all. These studies give in-depth insight into how users behave during search tasks which goes beyond what we model in this paper, although we too are interested in inferring user's reading and reformulation patterns.

Related Work The work by Guan et al (2013) on session search re-ranking based on query and impression term matching is a similar approach to ours, although we build a more complex model to capture user interactions and we do not perform document re-ranking. Another similar work is by Jiang et al (2014) who conduct a comprehensive user and eye tracking study to understand how users behave over the course of a session. Their work includes statistics on reformulation behavior and ranking metrics across queries in sessions and many of their results mirror our own. Both pieces of research can be seen as a specialization of our methodology (for instance focusing on a particular type of term source) that concerns a specific IR problem, whereas ours is a more general study on trying to understand reformulation behavior.

The work most similar to ours is the work by Liu et al (2011) on using terms from clicked snippets to aid in query recommendation. They recognize, as we do, that information needs persist through adjacent queries in search sessions but are difficult to define based purely on previous queries, and so use snippets as an additional term source. Unlike our methodology, they only use clicked snippets whereas we also incorporate terms from non-clicked snippets and documents, as well as the previous query. Where our work mainly diverges is that their objective is to locate terms that are useful for query recommendation, whereas our objective is to identify useful term sources for query reformulation (of which clicked snippets is one) under a number of conditions including clicks, rank and impression position.

The work in this paper differs from the literature in that: 1) our methodology is term-based rather than query or task-based 2) our methodology is derived from data rather than a user or eye-tracking study and 3) our model incorporates clicks and differentiates term sources such as snippets and documents as sources of reformulation terms.

Table 2 TREC 2011, 2012, 2013 and 2014 Session Track data overview.

	TREC Session Track			
	2011	2012	2013	2014
Number of topics	62	48	49	51
Number of sessions	76	98	116	1075
Number of impressions	280	297	471	3784
Number of $q_n \rightarrow q_{n+1}$ pairs	204	199	355	2709
Average number of terms in query	3.34	3.40	3.51	3.21

3 Analytical Setup

We conducted experiments using the TREC 2011, 2012, 2013 and 2014 Session Track data (Kanoulas et al, 2012, 2013), which contains search logs collected by the TREC organizers and grouped by session. Whilst particpants were given predefined topics to search over, the organizers recorded all of the displayed URLs, titles and snippets and also user interactions including clicks and document dwell time . The corpora used were the ClueWeb09[1] and ClueWeb12[2] datasets. Relevance judgments were also collected for documents related to each of the topics. See Table 2 for more detailed information about the datasets.

In comparison to commercial search logs, the TREC dataset is small. Moreover, the artificial setting in which the participants were recorded conducting session search makes analysis on its data difficult to apply to commercially used search systems. For the purpose of this study, the dataset is ideal in that it is the only publicly available search log that contains the rich impression data needed for our analysis, that is, clicks, dwell times and all ranked snippets and documents (not just clicked). Whilst our statistics may not exactly reflect those found in commercial logs, we believe our theoretical insights are transferable, can be readily reproduced, and our methodology applicable to any similarly rich dataset. Furthermore, our dataset proved large enough to give us statistically significant values in our experiments.

Sessions in the dataset are made up of a list of queries, each of which contains a ranking of M documents (typically $M = 10$), the snippets and titles of each document and a list of the documents that were clicked including their order and dwell time. In a session containing N queries, we refer to the n'th query as q_n and its query reformulation (if $n < N$) as q_{n+1}. We denote $\overrightarrow{q_n}$ as the term vector representation of the query (with term frequency as the term weights) and Q_n as the set of its terms t_n. Our analysis and experiments concern the changes between queries in a session, so we extract each pair of queries in a session $q_n \rightarrow q_{n+1}$ for $n = 1 \dots N - 1$.

An **Impression** refers to all of the search data related to a query such as the ranked list of documents and the clickthroughs. Elements of an impression include snippets (and their titles), clicks, dwell time and documents. In this dataset each session ends with a 'test' query intentionally containing no ranking, the original purpose being for researchers to create rankings for this query by utilizing the information in the session. In these cases we do not con-