Table 10 Change in value for metrics NDCG, NERR and MAP from $q_n \rightarrow q_{n+1}$ for each term action and term scenario. Bold values indicate a statistically significant difference in IR metric score ($p < 0.05$ under the Wilcoxon signed rank test).

		Scenario					
		1	2	4	5	6	8
Retained	NDCG	**0.078▲**	−0.205▼	**-0.120▼**	−0.009▼	−0.019▼	**-0.058▼**
	NERR	**0.080▲**	−0.216▼	**-0.146▼**	−0.004▼	−0.024▼	**-0.064▼**
	MAP	**0.004▲**	**-0.011▼**	0.010▲	0.001▲	−0.003▼	**-0.009▼**
Removed	NDCG	**0.058▲**	0.000	−0.069▼	0.006▲	0.006▲	**-0.148▼**
	NERR	**0.037▲**	−0.024▼	−0.063▼	−0.015▼	0.017▲	**-0.140▼**
	MAP	**0.005▲**	−0.004▼	0.000	**-0.003▼**	0.001▲	**-0.010▼**
Added	NDCG	−0.025▼	**-0.127▼**	−0.051▼	−0.023▼	−0.046▼	−0.091▼
	NERR	**-0.019▼**	**-0.123▼**	−0.082▼	−0.020▼	−0.052▼	**-0.073▼**
	MAP	**-0.007▼**	**-0.007▼**	0.002▲	0.000	−0.006▼	−0.006▼

in Table 9. Firstly, we find that all term actions in scenarios 1 and 5 (where terms are not found in clicked snippets or documents) are less likely to lead to a click. When clicked documents are taken into account (scenarios 2, 4, 6 and 8) the likelihood of a click in the next query is much higher. In particular, for scenarios 2 and 4 clicks were more likely after removing query terms then retaining them, a result mirroring what we found in Figure 5. Terms added from clicked documents and snippets were also highly likely to result in a click.

6.3.2 IR Metric Based Evaluation

Whilst clicks are important implicit signals of relevance, we can also make use of the TREC Session Track relevance judgments to evaluate the effectiveness of term actions. The majority of sessions in the dataset are linked to topics, for which documents have been assessed for relevance by human assessors on a scale from 0 to 4. For each impression in the data set we calculated the Normalized Expected Reciprocal Rank at rank position 10 (NERR), Normalized Discounted Cumulative Gain at position 10 (NDCG) and the Mean Average Precision (MAP). These metrics are widely used and well regarded in the IR community and the cutoff point at rank 10 was chosen in order to evaluate the quality of results in a typical impression. NERR is a metric that rewards displaying a highly relevant document at a high rank, NDCG measures the quality of the retrieved documents and their order and MAP balances precision and recall.

We measured the difference in scores for each of the metrics calculated for the rankings of q_n and q_{n+1} across each scenario and term action and our results are in Table 10. We see that when scenario 1 query terms are *retained* there is a significant improvement across all IR metrics, but otherwise for the other scenarios we see scores decreasing, significantly so for scenario 8. We also see a similar pattern for *removing* terms across all scenarios. Finally, for *added* terms the IR metrics decrease across all scenarios, significantly so for scenarios 1 and 2. These results indicate the existence of a general trend of decreasing IR score for adjacent queries, and we find that when we plot the scores across

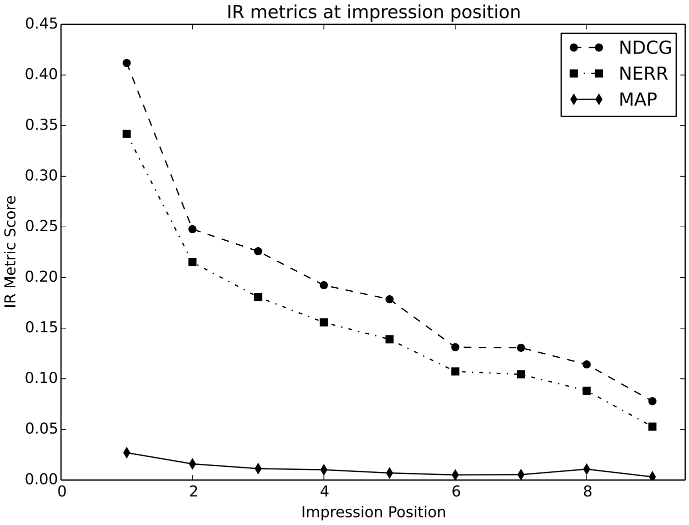

Fig. 6 NERR@10, NDCG@10 and MAP scores for user created rankings at each impression position.

impression positions (Figure 6) we confirm this negative trend. What we can take from these results is that when we come across an impression which doesn't contain query terms, the next query is likely to be an improvement (regardless of if the query term is retained or removed). Furthermore, in the converse scenario where query terms appear in all term sources, the next search ranking is likely to be worse.

We conclude on an interesting final result, where we see that when a term is added from a clicked document only (scenario 2), it leads to rankings with poorer IR scores. This is in spite of many of our findings that indicate that clicked documents are a rich source of added terms, that scenario 2 commonly occurs and that such reformulations lead to clicks 54.3% of the time. For example, the terms '`us government`' in Table 1 fall into scenario 2 for q_1 and are then added to q_2, whose ranking leads to a click and an improvement in NERR and NDCG, but not MAP, and then they are removed. Here, these terms represented a subtopic in the user's overall information need that was satisfied by their results before moving on. This result supports our argument that simply following the query reformulation behavior of users does not necessarily lead to improved search systems, but through understanding the interactions with our methodology we can make more informed inferences.

7 Discussion

Our novel methodology and term-based approach to understanding query reformulation leads to some interesting as well as expected results. We confirm that a user's query reformulation is largely made up of terms retained from their preceding query, with the remainder made up of a mix of terms discovered in the impression and externally sourced terms, although this can fluctuate throughout a session. However, we cannot expect to find all terms in a_{n+1} based on what's available in a query log because users introduce terms based on their own cognitive processes, memory, external context or when changing their information need. For instance, in the example in Table 1, the final query contains the term '`law`' that isn't found in any of the term sources in the previous impression and it's clear from the table that this query is a departure from the topic and pattern of the previous queries. In such cases, techniques such as behavioral modeling, ontologies, contextual retrieval and topic modeling could be used to predict new terms to add but this is beyond the scope of this work.

This work could be extended by further breaking down an impression into new term sources, such as snippet and document title or document components such as headers and paragraph text. Features such as rank and impression position or click order could be used to separate the current term sources and increase the number of scenarios. An n-gram model would require different similarity measures but would allow more accurate phrase matching and new term actions (such as phrase rearrangement, splitting etc.). Term sources from non-adjacent impressions could also help improve the overall model, and other implicit user measures (such as mouse tracking or reading level) could prove a good differentiator of term source similarity.

We are also aware that our analysis is restricted by the size and nature of the TREC session track data. An ideal analysis would be conducted over commercial query logs but these are not readily available. Also, the TREC data is flawed in that it has been compiled by researchers and doesn't strictly reflect an actual user interacting with a search engine. Nonetheless, the data does make up for these shortcomings with its rich meta-data, standardization and availability. Our inferences on query reformulation understanding are transferable to other areas of IR and our methodology can be readily applied to other datasets.

Our evidence suggests that user created query reformulations are not always successful and that it may be possible to generate viable reformulations (or suggestions) based on observing user feedback and classifying which scenarios terms belong to. Our intention is to use this research to build a query suggestion agent based on a Markov Decision Process (MDP) that incorporates our methodology, allowing us to create ranked lists of query suggestions using the retention, removal and addition term actions. By modeling the user's feedback using a Dynamic IR model (Jin et al, 2013), we can optimize the MDP over several projected queries in the session and predict the changing queries of the user, which will let us rank the most optimal query suggestion.