

Table 7 Overall average clicks, non-clicks and documents per impression and overall number of query and added term scenarios.

# ranked documents	10.5
# clicks	0.626
# non-clicks	9.87
# query term scenarios	7621
# added term scenarios	2981

Table 8 Scenario number definitions, occurrence % for query and added term scenarios and average number of ranked documents and clicks for each scenario.

Scenario	$t \in$			Query term scenarios			Added term scenarios		
	ncs	cs	cd	%	# Docs	# Clicks	%	# Docs	# Clicks
1	False	False	False	9.95	8.64	0.27	57.0	10.2	0.38
2	False	False	True	0.35	11.1	1.89	7.85	10.4	1.86
3	False	True	False	0.05	3.50	1.25	0.20	7.00	1.00
4	False	True	True	0.68	10.2	2.27	2.01	11.0	2.17
5	True	False	False	60.2	10.5	0.06	24.2	10.8	0.15
6	True	False	True	2.27	10.6	1.41	4.43	10.4	1.46
7	True	True	False	0.42	11.7	0.94	0.07	10.0	0.50
8	True	True	True	26.1	10.9	1.70	4.26	11.7	1.91

defines a scenario and we give a full index of scenario number definitions in Table 8.

In our previous analysis we were able to make inferences on terms based on which term source they originated from. With the expansion of 3 term sources to 8 scenarios we can now make more interesting observations. For instance, in the first query in our example (Table 1), the terms ‘gun’ and ‘control’ both belong to scenario 8 (they appear in non-clicked and clicked snippets and also clicked documents) and they are retained in the query reformulation. Conversely, the term ‘opinions’ is only found in non-clicked snippets (scenario 5) and is subsequently removed. An inference we can make here is that finding query terms in clicked snippets and documents is a strong indicator that the term will be kept, whereas query terms that only appear in non-clicked snippets are more likely to be removed.

We also assign added terms to the same scenarios in Table 8. Given that the purpose of our methodology is to understand when terms from the previous impression (including query) will be used in the reformulation, we appreciate that a real search system would not have access to added terms in order to assign them to scenarios. Nonetheless, by analyzing these terms in the same way as query terms, we gain insight into which circumstances a user is likely to add terms from the impression.

We extracted all query reformulation pairs from the dataset as before but this time did not include test queries (the final query in each session). Test queries do not contain rankings or relevance judgments, and thus are unsuitable for our evaluations in the next subsections. We assigned terms from q_n and a_{n+1} to each scenario and give an overview of our results in Tables 7 and 8.

In Table 8 we see that both sets of term scenarios are variably distributed. Scenario 5, which refers to the case where terms only appear in non-clicked snippets, is the most common scenario for query terms, comprising 60.2% of the data. For this scenario we find that the average number of clicks is 0.06, well below the overall average in Table 7. Thus, scenario 5 appears to be capturing the common case where users do not click on any results, hence no other term source matching occurs. Scenario 8 makes up a further 26.1% of cases and represents the situation where terms appear in all term sources. We would expect query terms to appear in snippets (either *ncs* or *cs*) and we find that this is the case 90% of the time. Interestingly, 9.95% of query terms do not appear in the impression at all.

We see a different distribution of scenarios for added terms, the most prominent being scenario 1 at 57%. This is the case where added terms cannot be found in the previous impression and mirrors the findings in Table 3. Scenario 5 is also common for added terms. The four scenarios with terms found in clicked documents (2, 4, 6 and 8) make up 18.6% of the scenarios, further evidence of clicked documents being a valuable source of added terms. There is a noticeable difference in occurrences between query and added terms in scenarios 2 and 4. Scenarios 3 and 7 rarely appear for both query and added terms, this can be explained by the fact that these are the cases where terms appear in clicked snippets but not clicked documents. Given that the snippet is derived from the document itself, this makes it unlikely for these scenarios to occur, and we ignore them in future analyses.

6.2 Term Actions

Query term scenarios fall into two term action categories, *retained* or *removed*. Figure 5 shows the proportion of query terms that are retained or removed from query reformulations for each scenario. Our first observation is that the two most common scenarios (5 and 8) lead to high term retention rates that are around the overall average term retention of 66%. This coincides with our earlier finding that users generally retain terms between adjacent queries, thus, the core terms are falling into these scenario numbers. For example, in Table 1 the query terms ‘gun control’ both belong to scenario 8 for the first 2 queries and are retained. For queries 3 and 4 (which are identical), they change to scenario 5 and are then subsequently removed in the next query.

Scenarios 2, 4 and 6, which capture instances of query terms appearing in clicked documents, occur infrequently for query terms and here seem to lead to the removal of terms. One inference is that terms appearing in clicked documents may be removed in lieu of the user satisfying that particular search intent. We also see low retention for query terms that are not found in the impression at all, potentially an indication that the term was not useful in helping the user’s search.

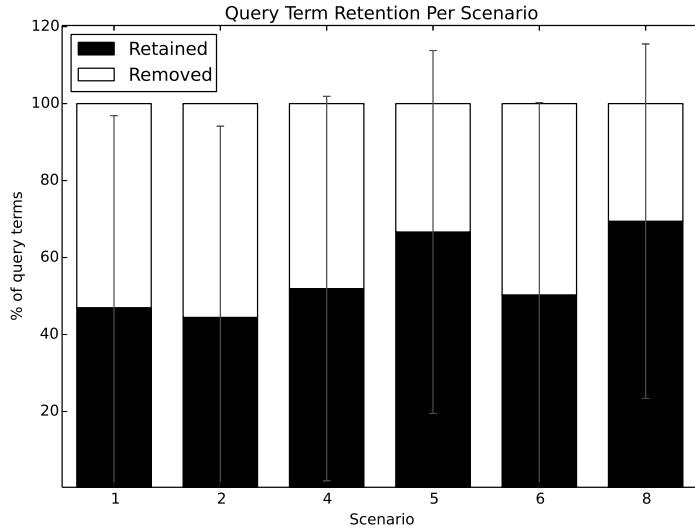


Fig. 5 Proportion of query terms that are retained or removed per term scenario

Table 9 Percentage of term scenarios and term actions that led to a click in the next query

Scenario	% (Term action → Click)		
	Retained	Removed	Added
1	22.5	29.1	26.3
2	25.0	53.3	54.3
4	59.3	68.0	63.3
5	24.5	22.1	27.5
6	41.4	52.3	59.1
8	52.3	49.1	64.6

6.3 Term Scenario Evaluation

So far we've sought to understand the term actions retention, removal and addition without explicitly evaluating whether or not they are beneficial. Simply determining which terms from queries and term sources are likely to appear in a reformulation, based on user behavior in search logs, does not necessarily mean that they will improve the search experience. These evaluations demonstrate that our methodology is able to differentiate scenarios which may lead to future clicks or improvements in IR metric scores.

6.3.1 Click Based Evaluation

Our first evaluation method involves observing whether the next impression in the session contains a click, an implicit measure of success and one tied to the user whose session we are analyzing. In this experiment, for each term scenario we measured the proportion of times each of the three term actions (retaining, removing or adding) led to a click in the next impression and give our results