Table 6 Average similarity of added terms with click-based variations of the snippet and document term sources and also the full preceding impression (i) and all previous impressions (h). Bold scores indicate a statistically significant ($p < 0.01$ under Welch's t-test) difference from non-clicked and 'All' variants of the term source.

Term Source	# terms	*Jaccard*	*Cosine*	$BM25$
All Snippets ($s(M)$)	50.4	0.00465	0.0175	0.688
Clicked Snippets (cs)	50.1	**0.00752**	**0.0289**	**1.100**
Non-Clicked Snippets (ncs)	50.5	0.00445	0.0167	0.660
All Documents (ad)	808.8	0.00131	0.0251	5.612
Clicked Documents (cd)	974.2	**0.00171**	**0.0398**	**8.207**
Non-Clicked Documents (ncd)	796.4	0.00128	0.0240	5.417
Impression (i)	8127.2	0.00067	0.0381	3.535
Historical (h)	19802.9	0.00052	0.0568	4.370

which we've already noted for its shift in query intent, we observe the added term '`center`' in the snippet at rank 3, which has the last (and only) click-through. This is in line with our findings on top ranked snippets in Table 4 and corroborates our last click hypothesis. Yet, at ranks 7 and 8 we see instances of the added term '`prevent`', suggesting that in this case the user examined snippets beyond the one that was clicked.

5.2 Term Sources

So far we have investigated the effect of impression and rank position on similarity and introduced clicks into our last experiment. Here we directly use clicks to further distinguish between the two distinct sources of added terms in an impression, snippets and documents. This allows us to split an impression into three term sources:

Non-Clicked Snippets (ncs) Snippets without a clickthrough.
Clicked Snippets (cs) Snippets with a clickthrough.
Clicked Documents (cd) Documents with a clickthrough

We note that the combination of nc and cs gives us all snippets in the impression i.e. $(\bigcup CS) \cup (\bigcup NCS) = S(M)$. We can now consider impression terms as belonging to one or more of the described term sources and start to evaluate how effective they are at providing added terms for query reformulations. Our reasoning for incorporating clicks into the term source definitions is that implicit user feedback is an indicator of the relevance of the terms contained in the source and the user's behavior at that point in the session.

Table 6 contains the results of our similarity analysis over different term sources and their variations with added terms. We compared clicked snippets and documents (cs and cd) with their non-clicked counterparts (ncs and ncd) and also against both combined ($s(M)$ and ad). We see in both cases statistically significant increases in similarity when considering clicks, a clear indicator that clicked snippets and documents are a source of terms used in query reformulations. Clicked documents score higher for the length normalized metrics *Cosine* and $BM25$ (the score is lower for the length biased *Jaccard* measure),

indicating the importance of clicked documents. We measured the similarity of non-clicked documents in order to provide comparison with clicked documents, but ultimately we do not consider them as a term source. This is because we cannot know if the user has been exposed to them during the session, although it is feasible that the user has encountered the document before or was satisfied by the non-clicked snippet itself.

We also measured the similarity with all terms found in the impression, where $I = S(M) + CD$ (not including the query). We find that differentiating an impression into click based term sources does lead to improved similarity scores. Taking this further, we also measured against historical impressions, i.e. all impression terms that occur earlier in the session up to and including the current impression $H_n = \bigcup_{j=1}^{n} I_j$, to test the assumption that users obtain terms not just from the preceding impression but also those encountered earlier. For instance, in our example in Table 1, the term '`current`' from q_3 is not found in the preceding impression for q_2, whereas it occurs 3 times in the snippet at rank 3 of q_1. We do see an increase in similarity scores over the historical impression terms and values that are comparable with the other term sources, suggesting that terms can be sourced from earlier in the session. In this work we define our term sources based only on the preceding impression, but using earlier impressions could prove an interesting extension.

5.3 Dwell Time

From Table 6 we see that clicked documents have substantially more terms than snippets. A central argument of our methodology is that users choose reformulation terms that they have been exposed to from term sources, hence, in order to come across terms in a long document, time must be spent reading it. Our dataset records the dwell time of each clicked document, which is an indicator of reading time.

We find that the average dwell time is 35.3 seconds before users return to the set of search results. This is similar to the 30 second threshold used in other IR research as a marker for a satisfactorily (SAT) clicked document (White and Drucker, 2007). SAT clicks are often used as a replacement for relevance judgments in the absence of human assessors, usually on large query logs. We find that a dwell time threshold of 30 seconds differentiates 40% of the clicked documents.

To test whether dwell time should be considered a feature in our methodology, we measured the similarity of clicked documents against added terms at a range of different dwell time thresholds. Figure 4 displays the results for *Cosine* similarity, the other measures reported similar findings. Whilst we do observe a slight increase in similarity with dwell time threshold, the results are too variable to be able to draw any conclusions. In particular, the SAT click threshold does not appear to offer any clear indicator of improvement. Our findings are supported by recent research that argues that this single value cannot capture the complexities of reading behavior and user satisfac-

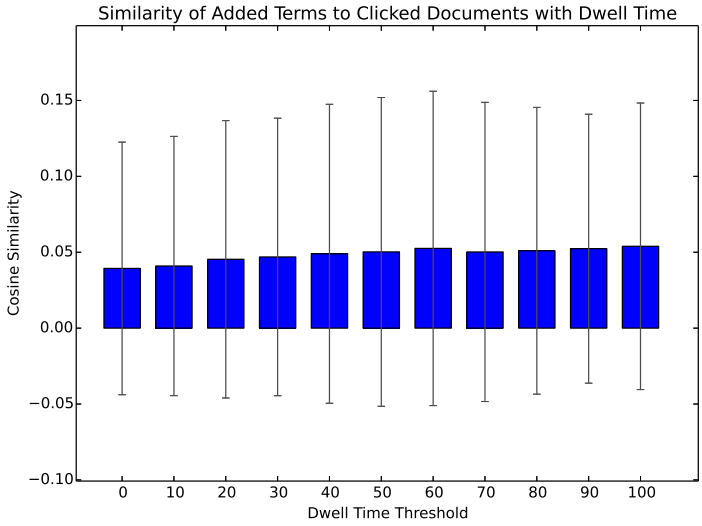

Fig. 4 Average *Cosine* similarity of added terms with clicked documents at different dwell time threshold levels.

tion (Kim et al, 2014). As such, we do not consider dwell time as a feature in our methodology and instead use all clicked documents as a term source collectively.

6 Term Scenario Analysis

Our term-based methodology has given us insight into the circumstances where terms are retained, removed or added to query reformulations. Use of the similarity measures has helped us define the three term sources, based on user interactions, that influence the terms added to the next query in a session. In this section, we extend our methodology to measure how effective query reformulations are under different circumstances. We do this by defining 8 user behavior scenarios based on the combination of term sources, which can help interpret our results and understand user motivations.

6.1 Query and Added Term Scenarios

We first focus on the query terms t_n and whether the term actions *retention* or *removal* are usually applied to them by the user. To expand on the limited information available to us on the terms in the query, we can look for occurrences of the term in the impression. More specifically, the three term sources *ncs*, *cs* and *cd*. t_n can belong to any combination of term sources, including all or none, giving 8 query **term scenarios**. Each combination of term source