What Can Task Teach Us About Query Reformulations?

Lynda Tamine, Jesus Lovon, Karen Pinel-Sauvagnat

▶ **To cite this version:**

Lynda Tamine, Jesus Lovon, Karen Pinel-Sauvagnat. What Can Task Teach Us About Query Reformulations?. European Conference on Information Retrieval - ECIR 2020, Apr 2020, Lisbon (on line), Portugal. pp.636–650, 10.1007/978-3-030-45439-5_42 . hal-02960357

HAL Id: hal-02960357
https://hal.science/hal-02960357

Submitted on 8 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

What Can Task Teach Us About Query Reformulations?

Lynda Tamine(✉), Jesús Lovón Melgarejo, and Karen Pinel-Sauvagnat

Université Paul Sabatier, IRIT, Toulouse, France
{tamine,sauvagnat}@irit.fr, jesus.lovon-melgarejo@univ-tlse3.fr

Abstract. A significant amount of prior research has been devoted to understanding query reformulations. The majority of these works rely on time-based sessions which are sequences of contiguous queries segmented using time threshold on users' activities. However, queries are generally issued by users having in mind a particular task, and time-based sessions unfortunately fail in revealing such tasks. In this paper, we are interested in revealing in which extent time-based sessions vs. task-based sessions represent significantly different background contexts to be used in the perspective of better understanding users' query reformulations. Using insights from large-scale search logs, our findings clearly show that task is an additional relevant search unit that helps better understanding user's query reformulation patterns and predicting the next user's query. The findings from our analyses provide potential implications for model design of task-based search engines.

Keywords: Tasks · Query reformulation · Query suggestion

1 Introduction

Query reformulation is a critical user behaviour in modern search engines and it is still addressed by a significant amount of research studies [10–12,17,23,26,33]. A salient behavioural facet that has been widely captured and analysed by those studies is query history. The latter is generally structured into "query sessions" which are sequences of queries submitted by a user while completing a search activity with a search system. In the literature review, there are many definitions of query sessions. The widely used definitions are the following [19,25]: (1) a *Time-based session*, also called physical session in [6], is a set of consecutive queries automatically delimited using a time-out threshold on user's activities. Time-gap values of 30 min and 90 min have been the most commonly used in previous research [4,6,9,19]; (2) a *Task-based session*, also called mission in [6], is a set of queries that are possibly neither consecutive nor within the same time-based session. The queries belong to related information needs that are driven by a goal-oriented search activity, called search *task* (eg., *job search* task). The latter could be achieved by subsets of consecutive related queries called logical sessions in [6] or subtasks in [9].

https://doi.org/10.1007/978-3-030-45439-5_42

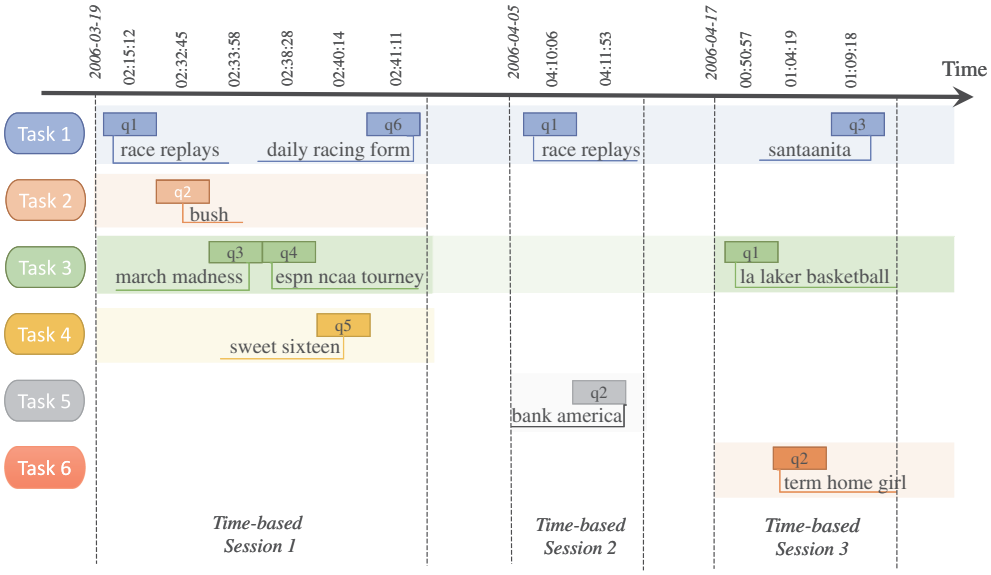


Fig. 1. Examples of time-based sessions and tasks, with the associated queries. Sample of the Webis-SMC-12 Search Corpus [6] for a given user.

Previous research [4,7,20,21] showed that: (1) users have a natural multi-tasking behaviour by intertwining different tasks during the same time-based session; and that (2) users possibly interleave the same task at different timestamps in the same time-based session or throughout multiple time-based sessions (ie., *multi-session tasks*). Such long-term tasks are acknowledged as being *complex tasks* [7,9]. Figure 1 shows a sample of 3 time-based search sessions extracted from the Webis-SMC-12 Search Corpus [6] for a single user. The sessions are manually annotated with tasks. As can be seen, 6 tasks (Task 1 - Task 6) are performed by the user during these 3 sessions. We can observe that all these sessions are *multi-tasking*, since they include queries that relate to multiple tasks (eg., Session 1 is multi-tasking since it includes queries that relate to Task 1, 2, 3 and 4). We can also see that Task 1 and Task 3 are *interleaved* within and across sessions (eg., Task 1 is interleaved within Session 1 and across Session 1, 2 and 3). Thus, Tasks 1 and 3 are *multi-session tasks*.

While it is well-known that time-based session detection methods fail in revealing tasks [6,19], most of previous research work has employed time-based sessions as the focal units of analysis for understanding query reformulations [10–12,26,33]. Other works rather studied users' query reformulations from the task perspective through user studies [15,17,29]. However, the authors analysed low-scale pre-designed search tasks conducted in controlled laboratory settings. In addition to their limited ability to observe natural search behaviour, there is a clear lack of comparability in search tasks across those studies.

To design support processes for task-based search systems, we argue that we need to: (1) fully understand how user's task performed in natural settings drives the query reformulations changes; and (2) gauge the level of similarity of these changes trends with those observed in time-based sessions. Our ultimate goal is to gain insights regarding the relevance of using user's tasks as the focal units of

search to both understand and predict query reformulations. With this in mind, we perform large-scale log analyses of users naturally engaged in tasks to examine query reformulations from both the time-based session vs. task-based session perspectives. Moreover, we show the role of the task characteristics in predicting the next user's query. Our findings clearly show that task is an additional relevant search unit that helps to better understand user's query reformulation patterns and to predict the next user's query.

2 Related Work

2.1 Query Reformulation Understanding

Query reformulation has been the focus of a large body of work. A high number of related taxonomies have been proposed [5,11,16]. To identify query reformulation patterns, most of the previous works used large-scale log analyses segmented into time-based sessions. Different time gaps have been used including 10–15 min [8], 30 min [4,19] and 90 min [6,9]. In a significant body of work, authors categorised the transitions made from one query to the subsequent queries through syntactic changes [11,12,23,26] and query semantic changes [10,12,33]. Syntactic changes include word substitution, removing, adding and keeping. The results highlighted that the query and its key terms evolve throughout the session regardless of the query position in the session. Moreover, such strategies are more likely to cause clicks on highly ranked documents. Further experiments on semantic query changes through generalisation vs. specialisation [10,12] showed that a trend exists toward going from generalisation to specialisation. This behavioural pattern represents a standard building-box strategy while specialisation occurs early in the session.

Another category of work rather employed lab user studies to understand how different task characteristics impact users' query reformulations [15,17,18,28,31,32]. The results mainly revealed that: (1) the domain knowledge of the task doer significantly impacts query term changes. For instance, Wildemuth [31] found that search tactics changed while performing the task as users' domain knowledge evolved; (2) the cognitive complexity and structure of the task (eg., simple, hierarchical, parallel) has a significant effect on users' query reformulation behavior. For instance, Liu et al. [17] found that specialisation in parallel tasks was significantly less frequent than in simple and hierarchical tasks.

A few work [4,22] used large-scale web search logs annotated with tasks to understand query reformulations. The findings in [4] were consistent with log-based studies [26] showing that page visits have significant influence on the vocabulary of subsequent queries. Odijk et al. [22] studied the differences in users' reformulation strategies within successful vs. unsuccessful tasks. Using a crowd-sourcing methodology, the authors showed that query specialisation through term adding is substantially more common in successful tasks than in unsuccessful tasks. It also appeared that actions such as formulating the same query than the previous one and reformulating completely a new query are rather relevant signals of unsuccessful tasks.

2.2 Contributions over Previous Work

We make several contributions over prior work. First, to the best of our knowledge, no previous study examined the differences in query reformulation strategies from the two perspectives of time-based sessions and task-based sessions viewed as background contexts. Insights gleaned from our data analysis have implications for designing task-based search systems. Second, although there has been intensive research on query reformulation, we provide a new insight into the variation of query reformulation strategies. The latter are analysed in relation with search episode size (*Short*, *Medium* and *Long*) and search stage (*Start*, *Middle* and *End*) from two different viewpoints (stream of query history and the search task progress). Third, building on the characterisation of search tasks, we provide insights on how considering task features might improve a supervised predictive model of query reformulations.

3 Analytical Set up

3.1 Datasets

This analysis is carried out using the freely available Webis-SMC-12 Search Corpus[1] [1,6] extracted from the 2006 AOL query log which is a very large collection of web queries. The released corpus comprises 8800 queries. We remove the repeated successive queries that were automatically generated following a click instead of a user's reformulation. We also remove all non-alphanumeric characters from the queries and apply a lowercasing. The cleaned data finally include 4734 queries submitted by 127 unique users. The query log is automatically segmented into time-based sessions using a time-gap threshold on users' activities. Since there is so far no agreement about the most accurate time-out threshold for detecting session boundaries [9,19], we consider the two widely used time-gap values between successive queries: 30 min as done in [4,19] and 90 min as done in [6,9]. We also use the provided manual annotations to segment the query log into task-based sessions. For care of simplicity, we subsequently refer to time-based session as "*Session*" and we refer to task-based session as "*Task*".

Table 1 presents the data collection statistics. One immediate observation is that the average number of queries in tasks (3.45) is higher than that of the sessions (eg., 2.04 in the 30 min-sessions) as reported in [9,19]. The total percentage of multi-tasking sessions is roughly 13% (resp. 16%) of the 30 min-session (resp. 90 min-session). Higher statistics (50%) were reported in [19]. However, we found that there are only 30.28% (resp. 31.27%) of the 30-min sessions (resp. 90-min sessions) that include only 1 task that is non interleaved throughout the user's search history. Thus, the 70% remaining sessions are either multi-tasking or include interleaved tasks that reoccur in multiple sessions. Similar statistics were observed in previous work (eg., 68% in [9]). Another interesting observation is that a high percentage of tasks (23.23%) are interleaved, which is roughly comparable to that of previous studies (eg., 17% in [14]), or spanned over multiple sessions (e.g, 27.09% of tasks spanned over multiple 30-min sessions).

[1] http://www.webis.de/research/corpora.

Table 1. The Webis search corpus statistics based on automatic segmentation of sessions (30 min, 90 min) and manual annotation of tasks.

	Sessions		Tasks
	30 min	90 min	
# of sessions/tasks	2318	2024	1373
Avg number of queries	2.04	2.34	3.45
Avg query length (#terms)	2.51	2.47	2.41
Multi-tasking sessions	12.87%	15.82%	-
Multi-session tasks	27.09%	25.42%	-
Interleaving tasks	-	-	23.23%

Table 2. Overview of query reformulation features.

Notation	Description	Measurement
$Sim(q_i, q_{i+1})$	Jaccard query pair similarity	$\frac{\lvert s(q_i) \cap s(q_{i+1}) \rvert}{\lvert s(q_i) \cup s(q_{i+1}) \rvert}$
$Rr(q_i, q_{i+1})$	Ratio of term-retention	$\frac{\lvert s(q_i) \cap s(q_{i+1}) \rvert}{\lvert s(q_i) \rvert}$
$Rm(q_i, q_{i+1})$	Ratio of term-removal	$\frac{\lvert s(q_i) - s(q_{i+1}) \rvert}{\lvert s(q_i) \rvert}$
$Ra(q_i, q_{i+1})$	Ratio of term-adding	$\frac{\lvert s(q_{i+1}) - s(q_i) \rvert}{\lvert s(q_{i+1}) \rvert}$

3.2 Query Reformulation Features

To study query reformulations, we consider the three usual categories of syntactic changes [11,13,26] between successive query pairs (q_i, q_{i+1}) composed of $s(q_i)$ and $s(q_{i+1})$ term sets respectively: (1) query term-retention Rr; (2) query term-removal Rm acts as search generalisation [12,13]; and (3) query term-adding Ra acts as search specialisation [12,13]. For each query pair, we compute the similarity and the query reformulation features presented in Table 2, both at the sessions and tasks levels (Sect. 5).

4 Query Characteristics

4.1 Query Length

Here, our objective is twofold: (1) we investigate how query length (ie., # query terms) varies across the search stages within sessions and tasks of different sizes (ie., # queries); and (2) we examine in what extent the trends of query length changes observed within tasks are similar to those observed within sessions.

To make direct comparisons of trends between sessions and tasks with different sizes in a fair way, we first statistically partition the search sessions and tasks into three balanced categories (*Short*, *Medium* and *Long*). To do so, we compute the cumulative distribution function (CDF) of session size values for the 30-min and the 90-min sessions, as well as the CDF of task size values in

Table 3. Classification of sessions and tasks regarding the number of related queries. If applicable, query positions boundaries to delimit the search stages in sessions and tasks of different sizes.

	Short	Medium		Long		
Sessions (30 min, 90 min)	1	2		≥3		
		Query position boundary		*Query position boundary*		
		Start	Middle	Start	Middle	End
		1	2	1–2	3	≥4
Tasks	1	2		≥3		
		Query position boundary		*Query position boundary*		
		Start	Middle	Start	Middle	End
		1	2	1–3	4–8	≥9

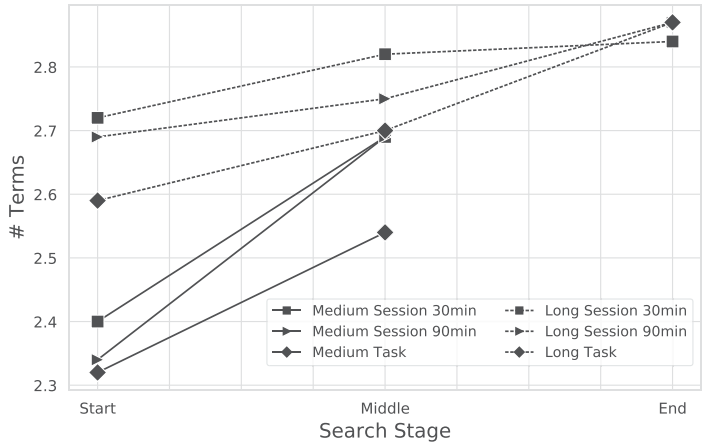
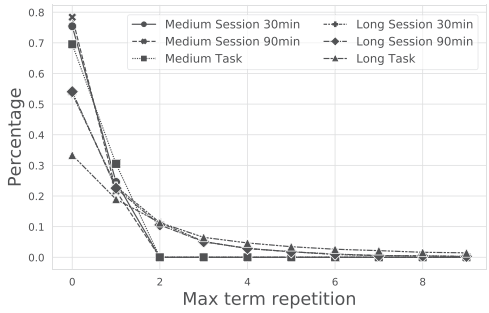


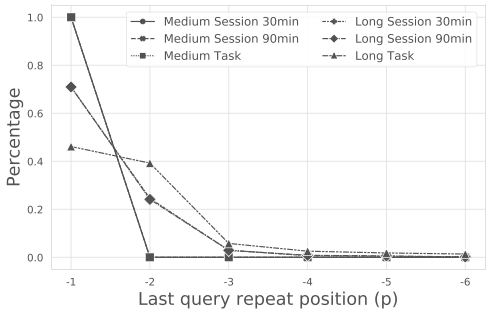
Fig. 2. Average query length variation along sessions vs. tasks of different sizes.

relation with the number of included queries. Then, we compute the CDF of the search stage values in relation with the query position boundary (*Start*, *Middle* and *End*) along each size-based category of sessions vs. tasks. Since short sessions and tasks only contain 1 query and consequently do not contain query reformulations, we do not distinguish between the search stages nor consider this category of sessions and tasks in the remainder of the paper. Table 3 shows the statistics of the search stages (*Start, Middle, End*) with respect to *Medium* and *Long* sessions and tasks.

Based on those categorisations, Fig. 2 shows the variation of the query length limit within each category of sessions and tasks and along the different search stages. We can see two clear trends. First, queries in both longer sessions and longer tasks generally tend to contain more terms (2.60–2.87 vs. 2.41–2.51 in average). This trend remains along all the different search stages. Regarding sessions, previous studies [2] have also shown similar trends in log-based data. Regarding tasks, our results suggest that long tasks require to issue more search terms. One could argue that long tasks, that more likely involve complex



(a) The percentage of sessions vs. tasks of different sizes with corresponding maximum term repeat.



(b) The percentage of queries containing the same term to a previous query over different positions.

Fig. 3. Term repetition trends over sessions vs. tasks.

information needs, lead users to formulate more informative queries. We also relate this observation with previous findings [2] showing that increased success is associated with longer queries, particularly in complex search tasks. Second we can surprisingly see that in general, queries observed within sessions whatever their sizes, are slightly longer in average than queries issued within tasks of the same category except at the end of the search stage. By cross-linking with the CDF results presented in Table 3, we expect that this observation particularly relates to long sessions. One possible explanation is that since long sessions are more likely to be multi-tasking (eg., there are 1.57 task in average in the long 90-min sessions vs. 1.29 in the 30-min sessions), the average query length is particularly increased within sessions that include queries at late search stages of the associated tasks (*Middle*, *End*).

4.2 Query Term Repeat

Inspired by [13], we examine query term frequency along the search with respect to session vs. task search context. In contrast to [13], our underlying intent here is rather to learn more about the impact of search context (ie., session vs. task) on the level of query term reuse. For a query q_i belonging to session S and task T and not submitted at the beginning (ie., $i > 1$), we compute the frequency of each of its terms from the previous queries within the same session q_j^S (resp. same task q_j^T), $j = 1..i-1$. Then, we take the maximal value Tr as "maximum term repeat" for query q_i if the latter contains at least one term used Tr times in previous queries.

Figure 3a plots the average "maximum term repeat values" for all the queries within all the sessions and tasks ranged by size (*Short*, *Medium* and *Long*). We can see that the term repeat trend across sessions is similar to that reported in [13]. By comparing between the term repeat trends in sessions and tasks, we clearly observe that there are less reformulated queries that do not share any identical terms with the previous queries in tasks (eg., 70% of medium tasks) in comparison to sessions (eg., 75–78% of medium sessions). Interestingly, we can see that the difference is particularly higher in the case of long tasks and long

sessions (33% vs. 53–54%). However, we can notice that even if the percentage of queries sharing an increased number of terms with previous queries decreases for both medium sessions and medium tasks, the difference is reversed between long sessions and long tasks. It is more likely that query terms are renewed during long tasks which could be explained by shifts in information needs related to the same driving long-term task.

Figure 3b shows the percentage of reformulated queries for which each reused term occurs at the first time at a given position within sequences from length 1 to 6. It appears that the sources of reused query terms in both tasks and sessions are limited to the two previous queries. More particularly, while we find terms used in the previous query in all (100%) of the reformulated queries in medium sessions and medium tasks, it is more likely to observe reformulated queries containing terms from the two previous queries in long sessions than in long tasks (71% of sessions vs. 46% of tasks). To sum up, the context used for driving query actions is limited to the two previous queries even for long sessions and tasks, with however, a lower level of term reuse in long tasks.

5 Query Reformulation

5.1 User Actions

Given each query q_i belonging to session S (resp. task T), Table 4 gives the query reformulation feature values (See Table 2) for both *Medium* (M) and *Long* (L) sessions and tasks and are computed over: (1) the short-term context (SC), by considering the query reformulation pair observed within the same session S (resp. task T) $(q_i, q_{i+1})^S$ (resp. $(q_i, q_{i+1})^T$), $i \geq 1$; and (2) the long-term context (LC), by considering the set of successive query reformulation pairs within the same session S (resp. task T), $(q_k, q_{k+1})^S$ (resp. $(q_k, q_{k+1})^T$), $1 \leq k \leq i$. Significance of the differences between the "Within Session" scenario and the "Within Task" scenario considering either the short-term context (SC) or the long-term context (LC) is computed using the non-paired student t-test. We can see from Table 4 that for the whole set of search actions (ie., term-retention *Rr*, term-removal *Rm* and term-adding *Ra*) and similarity values (ie., *Avg Sim*), most of the differences between task-based and session-based scenarios are highlighted as significant. More particularly, we can make two key observations: (1) successive queries in both medium and long tasks are significantly more similar (*Avg Sim* of 0.27 and 0.25 respectively) than they are in medium and long sessions for both time-out thresholds (*Avg Sim* of 0.20–0.23) with higher ratios of term-retention (34% vs. 25–29%); and (2) the query history along long tasks exhibits a higher topical cohesion (*Avg Sim* of 0.24) than it does in long sessions (*Avg Sim* of 0.18–0.20) with a higher ratio of term-retention (30% vs. 23–26%) and a lower ratio of term-adding (70% vs. 74–77%) for tasks. All these results are consistent with those obtained through the analysis of query term repeat (Sect. 4.2). They suggest that longer tasks more likely include topically and lexically closer information needs that might drive subtasks in comparison with long sessions.

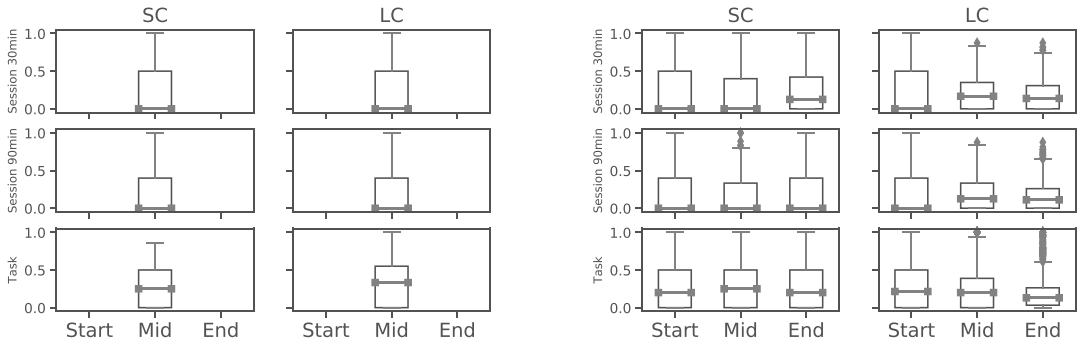
Unlikely, the latter might include multiple and topically different information needs that belong to distinct tasks.

5.2 Similarity Analysis over the Search Progress

To better understand the changes trends along the search, we also examine (Fig. 4) the query reformulation similarities at different stages of the search sessions vs. tasks by considering both short-term context (SC) and long-term context (LC). We can make from Fig. 4 two important observations: (1) successive query reformulations within tasks are clearly more similar ($m = 0.25$, $sd = 0.27$, $avg = 0.27$) at the different search stages than they are within both the 30-min and 90-min sessions (eg., $m = 0.0$, $sd = 0.27$, $avg = 0.23$ for the 30-min sessions) regardless of their sizes; and (2) the overall similarity of query reformulations observed over the search history in both long sessions and long tasks tends to decrease along the search (eg., decrease from $m = 0.21$ to $m = 0.13$ for tasks). These results indicate that the queries tend to be lexically dissimilar while the search evolves. This observation might be explained by different reasons

Table 4. Reformulation and similarity feature values in sessions vs. tasks. Significant differences ($p < 0.01$) of the "Within Session" scenario in comparison to the "Within Task" scenario are highlighted using a star '*'.

Features	Within Task (Baseline)				Within Session							
	SC		LC		30 min				90 min			
					SC		LC		SC		LC	
	M	L	M	L	M	L	M	L	M	L	M	L
Avg Sim.	0.27	0.25	0.35	0.24	0.23	0.22*	0.23*	0.20*	0.20*	0.21*	0.20*	0.18*
Rr	0.34	0.34	0.41	0.30	0.29	0.29*	0.29*	0.26*	0.25*	0.28*	0.25*	0.23*
Rm	0.61	0.63	0.54	0.69	0.66	0.68*	0.66*	0.72*	0.71*	0.70*	0.71*	0.75*
Ra	0.66	0.66	0.59	0.70	0.71	0.71*	0.71*	0.74*	0.75*	0.72*	0.75*	0.77*



(a) Medium sessions and medium tasks. **(b)** Long sessions and long tasks

Fig. 4. Plots of query reformulation similarities along different search stages of sessions and tasks of different sizes.

depending on the context used (session vs. task) to make the observation. As outlined earlier through query length analysis (Sect. 4.1), sessions might include different ongoing tasks that lead to formulate lexically distinct queries. Unlikely, tasks might include different ongoing related subtasks. However, queries are still overall more similar ($m = 0.13$, $sd = 0.23$, $avg = 0.20$) across the search stages in long tasks than they are in long sessions ($m = 0.11$, $sd = 0.17$, $avg = 0.16$), particularly at the end of the search stage. This observation might be related to the better cohesiveness of tasks with increased number of queries since, unlike sessions, they are goal-oriented.

5.3 Summary

Through the analyses presented in the previous sections, we have shown that there are significant differences in query reformulation patterns depending potentially on the context used (session or task) to make the observations. The results also indicate that time threshold value used to segment the sessions has no impact on the differences trends. In general, the most significant differences are observed regarding long tasks. Informed by these findings, we show in the final contribution of this paper the potential of the task features studied in Sects. 4 and 5 for enhancing the performance of a query reformulation predictive model.

6 Predicting Query Reformulation Using Task Context

Given a session $S = \{q_1, q_2, \ldots, q_{M-1}, q_M\}$, we aim to predict for each query sequence $S_k \subset S, S_k = \{q_1, q_2 \ldots, q_{k-1}, q_k\}$, $1 < k < M$, the target query q_k given the context C_{q_k} defined by queries $\{q_1, q_2 \ldots, q_{k-1})$, where q_{k-1} is the anchor query.

6.1 Experimental Setting

Evaluation Protocol. As usually done in previous work for query auto-completion [13] and next query prediction [3,24,27], we adopt a train-test methodology. We first sort the 30 min-sessions time-wise and partition them into two parts. We use the first 60 day-data for training the predictive model and the remaining 30 days for testing. We use 718 sessions (including 2418 queries) which represent 70% of the dataset as our training set, and 300 sessions (including 998 queries) which represent 30% of the dataset as our testing set. To enable the evaluation of the learning approach, we first produce a set of ground truth suggestions for each test query. To do so, we follow a standard procedure [3,13,27]: for each session in the training-test sets, we select as the candidate set, the top-20 queries q_k that follows each anchor query q_{k-1}, ranked by query frequency. To assess the contributions of the task context features in predicting the next user's query, we use the *Baseline Ranker*, a competitive learning to rank query suggestion model that relies on contextual features [3,27].

Model Training. We design the task-aware *Baseline Ranker* which we refer to as *TaskRanker*. For training purpose, we first generate from the 718 training sessions, 1395 task-based query sequences that are built with respect to the task labels provided in the Webis-SMC-12 Search Corpus. We remove the task-based query sequences with only 1 query candidate. For instance, using task labels provided in Fig. 1, we built and then select from *Session 1* the task-based query sequences $\{q1, q6\}; \{q3, q4\}$ with respectively $q6$ and $q4$ as the ground truth queries. Besides, to guarantee the candidate set includes the target query, we remove the task-based query sequences whose ground truth is not included in the associated candidate sets. After filtering, we obtain 215 cleaned task-based query sequences used for training the *TaskRanker* model. Similarly to [3,27], we use the state-of-the-art boosted regression tree ranking algorithm LamdaMART as our supervised ranker. We tune the LamdaMART model with parameters of 500 decision trees across all experiments. We use 2 sets of features (30 in total): (1) 10 features related to the analyses conducted in previous sections of the paper (Sects. 4, 5). We use the *user-action related features* including ratios of term-retention (Rr), term-adding (Ra), term-removal (Rm), and term-repeat (Tr), that are measured using both the short-term (SC) and long-term (LC) contexts. We also use *query-similarity related features* ($Avg\ Sim$) based on the similarity of the target query q_k with short-term context SC (anchor query q_{k-1}) and long-term context LC (with the previous queries in C_{q_k}); (2) 20 features that are similar to those previously used for a learning to rank suggestion model, and described in detail in [3,27]. This set of features includes (a) pairwise and suggestion features based on target query characteristics and anchor query characteristics including length and frequency in the dataset; (b) contextual features that include n-gram similarity values between the suggestion and the 10 most recent queries. Note that we extended the *Baseline Ranker* released by Sordoni et al. [27][2].

Baselines and Evaluation Metric. We use the conventional models widely used in the literature [3,13,27] namely the *Most Popular Suggestion (MPS)*, and the traditional *Baseline Ranker* which we refer to as *SessionRanker*. The MPS relies on query frequency to rank candidates. Unlike the *TaskRanker*, the *SessionRanker* is trained on session-based query sequences that are built from the same subset of the 718 training sessions. For instance, we built from *Session 1* presented in Fig. 1, the session-based query sequences $\{q1, q2\}; \{q1, q2, q3\}; \{q1, q2, q3, q4\}; \{q1, q2, q3, q4, q5\}; \{q1, q2, q3, q4, q5, q6\}$ with respectively $q2$, $q3$, $q4$, $q5$ and $q6$ as the ground truth queries. We obtain 1700 session-based query sequences that are then cleaned, similarly to the *TaskRanker* by removing query sequences with only 1 query candidate and those with ground truth not included in the associated candidate sets. Finally, the *SessionRanker* has been trained on 302 cleaned session-based query sequences.

[2] https://github.com/sordonia/hred-qs.

Similarly to the *TaskRanker*, we use the same sets of features (30 in total) learned here at the session level, and we tune it using the LamdaMART model. We use the *Mean Reciprocal Rank* (*MRR*) which is the commonly used metric for evaluating next query prediction models [3,24,27]. The MRR performance of the *TaskRanker* and the baselines is measured using the same test subset that includes 150 cleaned session-based query sequences built up on the subset of 698 session-based query sequences generated from the 300 test sessions. The task annotations of the testing test are ignored.

6.2 Prediction Results

Table 5 shows the MRR performance for the *TaskRanker* and the baselines. The *TaskRanker* achieves an improvement of +152.8% with respect to the MPS model and an improvement of +10.2% with respect to the *SessionRanker* model. The differences in MRR are statistically significant by the t-test ($p < 0.01$). It has been shown in previous work [3,27] that session size has an impact on the performance of context-aware next query prediction models. Thus, we report in Fig. 5 separate MRR results for each of the *Medium* (2 queries) and the *Long* sessions ($\geq$3 queries) studied in our analyses (Sects. 4 and 5). As can be seen, the task-based contextual features particularly help predicting the next query in long sessions ($+14,1\%$ in comparison to the *SessionRanker*, $p = 7 \times 10^{-3}$). Prediction performance for *Medium* sessions is slightly but not significantly lower ($-1,3\%$ in comparison to the *SessionRanker*, $p = 0.65$). This result can be expected from

Table 5. Next query prediction performance. All improvements are significant by the t-test ($p < 0.01$).

Model	MRR	Improvement
MPS	0.3677	+152.8%
SessionRanker	0.8433	+10.2%
TaskRanker	**0.9296**	–

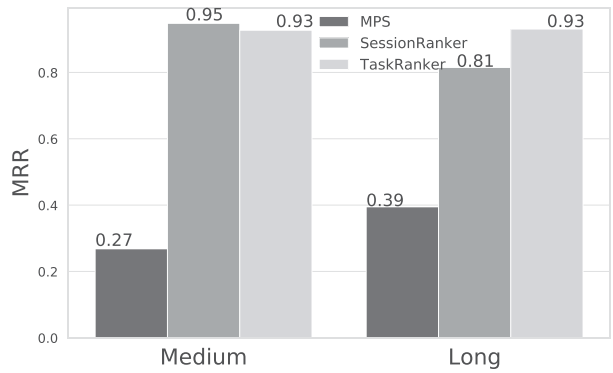


Fig. 5. Performance of *TaskRanker* compared to the baseline models on sessions with different sizes.

the findings risen from our analyses, since *Long* sessions include queries related to 89.9% of *Long* tasks whose cohesive contexts enable more accurate predictions of user's future search intent.

7 Conclusion and Implications

Better understanding user's query reformulations is important for designing task completion engines. Through the analysis of large-scale query logs annotated with task labels, we have revealed significant differences in the query changes trends along the search depending on the retrospective context used, either session or task. We found that queries are even longer in longer tasks with however a lower level of term reuse in tasks than in sessions. In addition, terms are particularly renewed in long tasks indicating clear shifts in information needs. Using lexical similarity measures, we have also shown that the query reformulations exhibit a clearer cohesiveness within tasks than within sessions along the different search stages, with however a decreasing level of similarity. Finally, we provided insights on the usefulness of task features to enhance the user's next query prediction accuracy. Given the crucial lack of query logs with annotated tasks, we acknowledge that the predictive model has been trained and tested with limited amount of data. However, the features used are based on the analysis performed on a large-scale data provided in the Webis corpus. Thus, we believe that the trend of our results would remain reliable.

There are several promising research directions for future work. Firstly, evidence related to the characterization of tasks through query length variation and query reformulation similarities along the search, presented in Sects. 4 and 5, may benefit research on automatic task boundary detection. In Sect. 6, we showed that learning from query streams annotated with tasks helps the query suggestion process particularly for long-term tasks. It will be interesting to design a predictive model of *query trails* associated with subtasks, by analogy to search trails [30]. This might help users in completing complex tasks by issuing fewer queries. This would decrease the likeliness of search struggling as shown in previous work [22].

References

1. Webis corpus archive. https://zenodo.org/record/3265962#.Xc8HoS2ZPOQ. https://doi.org/10.5281/zenodo.3265962
2. Agapie, E., Golovchinsky, G., Qvarfordt, P.: Leading people to longer queries. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2013, pp. 3019–3022 (2013)
3. Dehghani, M., Rothe, S., Alfonseca, E., Fleury, P.: Learning to attend, copy, and generate for session-based query suggestion. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, pp. 1747–1756 (2017)

4. Eickhoff, C., Teevan, J., White, R., Dumais, S.: Lessons from the journey: a query log analysis of within-session learning. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, WSDM 2014, pp. 223–232 (2014)
5. Guo, J., Xu, G., Li, H., Cheng, X.: A unified and discriminative model for query refinement. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp. 379–386 (2008)
6. Hagen, M., Gomoll, J., Beyer, A., Stein, B.: From search session detection to search mission detection. In: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, OAIR 2013, pp. 85–92 (2013)
7. Hassan Awadallah, A., White, R.W., Pantel, P., Dumais, S.T., Wang, Y.M.: Supporting complex search tasks. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, pp. 829–838 (2014)
8. He, D., Göker, A.: Detecting session boundaries from web user logs. In: In Proceedings of of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research, pp. 57–66 (2000)
9. He, J., Yilmaz, E.: User behaviour and task characteristics: a field study of daily information behaviour. In: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, pp. 67–76 (2017)
10. He, Y., Tang, J., Ouyang, H., Kang, C., Yin, D., Chang, Y.: Learning to rewrite queries. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM 2016, pp. 1443–1452 (2016)
11. Huang, J., Efthimiadis, E.N.: Analyzing and evaluating query reformulation strategies in web search logs. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 77–86 (2009)
12. Jansen, B.J., Booth, D.L., Spink, A.: Patterns of query reformulation during web searching. J. Am. Soc. Inf. Sci. Technol. **60**(7), 1358–1371 (2009)
13. Jiang, J.Y., Ke, Y.Y., Chien, P.Y., Cheng, P.J.: Learning user reformulation behavior for query auto-completion. In: Proceedings of the 37th International ACM SIGIR Conference on Research & #38; Development in Information Retrieval, SIGIR 2014, pp. 445–454 (2014)
14. Jones, R., Klinkner, K.L.: Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, pp. 699–708 (2008)
15. Kinley, K., Tjondronegoro, D.W., Partridge, H.L., Edwards, S.L.: Human-computer interaction: the impact of users' cognitive styles on query reformulation behaviour during web searching. In: Australasian Conference on Computer-Human Interaction (OZCHI 2012), Melbourne, Vic, August 2012. https://doi.org/10.1145/2414536.2414586
16. Lau, T., Horvitz, E.: Patterns of search: analyzing and modeling web query refinement. In: Kay, J. (ed.) UM99 User Modeling. CICMS, vol. 407, pp. 119–128. Springer, Vienna (1999). https://doi.org/10.1007/978-3-7091-2490-1_12
17. Liu, C., Gwizdka, J., Liu, J., Xu, T., Belkin, N.J.: Analysis and evaluation of query reformulations in different task types. In: Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem, ASIS&T 2010, vol. 47, pp. 17:1–17:10 (2010)

18. Lu, K., Joo, S., Lee, T., Hu, R.: Factors that influence query reformulations and search performance in health information retrieval: a multilevel modeling approach. J. Assoc. Inf. Sci. Technol. **68**(8), 1886–1898 (2017)
19. Lucchese, C., Orlando, S., Perego, R., Silvestri, F., Tolomei, G.: Identifying task-based sessions in search engine query logs. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM 2011, pp. 277–286 (2011)
20. Mehrotra, R., Bhattacharya, P., Yilmaz, E.: Characterizing users' multi-tasking behavior in web search. In: Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval, CHIIR 2016, pp. 297–300 (2016)
21. Mehrotra, R., Bhattacharya, P., Yilmaz, E.: Uncovering task based behavioral heterogeneities in online search behavior. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2016, pp. 1049–1052. ACM, New York (2016). https://doi.org/10.1145/2911451.2914755
22. Odijk, D., White, R.W., Hassan Awadallah, A., Dumais, S.T.: Struggling and success in web search. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM 2015, pp. 1551–1560 (2015)
23. Rieh, S.Y., Xie, H.I.: Analysis of multiple query reformulations on the web: the interactive information retrieval context. Inf. Process. Manag. **42**(3), 751–768 (2006)
24. Santos, R.L.T., Macdonald, C., Ounis, I.: Learning to rank query suggestions for adhoc and diversity search. Inf. Retrieval **16**(4), 429–451 (2013). https://doi.org/10.1007/s10791-012-9211-2
25. Silverstein, C., Marais, H., Henzinger, M., Moricz, M.: Analysis of a very large web search engine query log. SIGIR Forum **33**(1), 6–12 (1999)
26. Sloan, M., Yang, H., Wang, J.: A term-based methodology for query reformulation understanding. Inf. Retrieval J. **18**(2), 145–165 (2015). https://doi.org/10.1007/s10791-015-9251-5
27. Sordoni, A., Bengio, Y., Vahabi, H., Lioma, C., Grue Simonsen, J., Nie, J.Y.: A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM 2015, pp. 553–562 (2015)
28. Tamine, L., Chouquet, C.: On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings. Inf. Process. Manag. **53**(2), 332–350 (2017)
29. Vakkari, P.: A theory of the task-based information retrieval. J. Doc. **57**, 44–60 (2001)
30. White, R.W., Huang, J.: Assessing the scenic route: measuring the value of search trails in web logs. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, pp. 587–594 (2010)
31. Wildemuth, B.M.: The effects of domain knowledge on search tactic formulation. JASIST **55**, 246–258 (2004)
32. Wildemuth, B.M., Kelly, D., Boettcher, E., Moore, E., Dimitrova, G.: Examining the impact of domain and cognitive complexity on query formulation and reformulation. Inf. Process. Manag. **54**(3), 433–450 (2018)
33. Ãzmutlu, H., Cavdur, F.: Application of automatic topic identification on excite web search engine data logs. Inf. Process. Manag. **41**, 1243–1262 (2005). https://doi.org/10.1016/j.ipm.2004.04.018