Towards a Better Understanding of Query Reformulation Behavior in Web Search

Jia Chen[†], Jiaxin Mao[‡], Yiqun Liu[†★], Fan Zhang[†], Min Zhang[†], Shaoping Ma[†]
† Department of Computer Science and Technology, Institute for Artificial Intelligence
Beijing National Research Center for Information Science and Technology
Tsinghua University, Beijing 100084, China
‡ Gaoling School of Artificial Intelligence
Renmin University of China, Beijing 100872, China
yiqunliu@tsinghua.edu.cn

ABSTRACT

As queries submitted by users directly affect search experiences, how to organize queries has always been a research focus in Web search studies. While search request becomes complex and exploratory, many search sessions contain more than a single query thus reformulation becomes a necessity. To help users better formulate their queries in these complex search tasks, modern search engines usually provide a series of reformulation entries on search engine result pages (SERPs), i.e., query suggestions and related entities. However, few existing work have thoroughly studied why and how users perform query reformulations in these heterogeneous interfaces. Therefore, whether search engines provide sufficient assistance for users in reformulating queries remains under-investigated. To shed light on this research question, we conducted a field study to analyze fine-grained user reformulation behaviors including reformulation type, entry, reason, and the inspiration source with various search intents. Different from existing efforts that rely on external assessors to make judgments, in the field study we collect both implicit behavior signals and explicit user feedback information. Analysis results demonstrate that query reformulation behavior in Web search varies with the type of search tasks. We also found that the current query suggestion/related query recommendations provided by search engines do not offer enough help for users in complex search tasks. Based on the findings in our field study, we design a supervised learning framework to predict: 1) the reason behind each query reformulation, and 2) how users organize the reformulated query, both of which are novel challenges in this domain. This work provides insight into complex query reformulation behavior in Web search as well as the guidance for designing better query suggestion techniques in search engines.

∗ This work is supported by the National Key Research and Development Program of China (2018YFC0831700), Natural Science Foundation of China (Grant No. 61732008, 61532011), Beijing Academy of Artificial Intelligence (BAAI) and Tsinghua University Guoqiang Research Institute.
★Corresponding author

WWW '21, April 19-23, 2021, Ljubljana, Slovenia

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

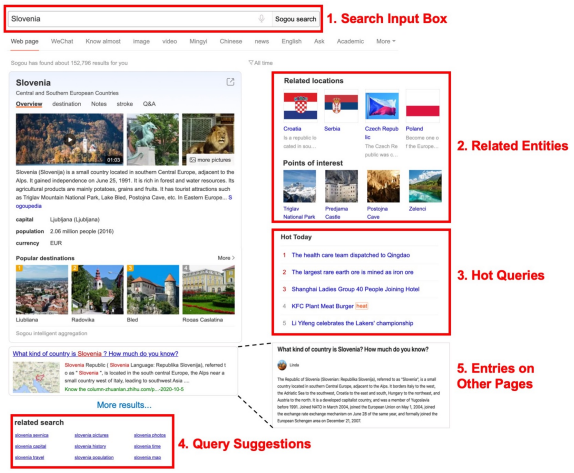


Figure 1: Typical entries provided for query reformulation in a popular commercial search engine.

KEYWORDS

User Behavior Analysis, Query Reformulation, Behavior Prediction

ACM Reference Format:
Jia Chen[†], Jiaxin Mao[‡], Yiqun Liu[†★], Fan Zhang[†], Min Zhang[†], Shaoping Ma[†]. 2021. Towards a Better Understanding of Query Reformulation Behavior in Web Search. In *Proceedings of The Web Conference 2021 (WWW '21).* ACM, New York, NY, USA, 12 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

In complex search scenarios, users usually strive for useful information by reformulating their queries in multiple search rounds. As queries submitted by users directly affect their search experiences, query reformulation has always been a bottleneck issue in Web search. It is therefore of vital importance to provide query reformulation supports in SERPs.

To help search engines better fulfill users' information needs, a large body of research has focused on designing better frameworks for query suggestion or query auto-completion [7, 22, 35]. However, these data-driven methods only relied on coarse-grained representations of the users' previous knowledge within the session to fit the observational data, i.e., predicting the next query [9]. More in-depth investigations should be conducted on both the motivation and the pattern of user reformulations. Based on search engine logs, existing

efforts have analyzed users' reformulating behaviors [12, 14, 21, 31]. As search logs are noisy and only contain implicit feedback from users, some researchers conducted user studies to collect richer data in a more controlled environment [9, 19, 27]. However, few existing studies on query reformulation have taken a further step in understanding user intents behind the query reformulation. Therefore, besides predicting the content of the next query, we need to further investigate why and how users reformulate a query for the optimization of search engines.

Moreover, existing studies were mainly designed for text-only SERPs. As modern search engines usually provide a series of entries (or interfaces) for users to better reformulate their queries (see Figure 1 as an example), user behavior may also be affected by these entries. However, to our best knowledge, most existing studies ignore user interactions with these heterogeneous entries. How users will reformulate their queries with these supports from search engines is still under-investigated. Therefore, understanding how users utilize these interfaces is necessary for further improving them.

So in this study, we go beyond predicting the next queries and thoroughly investigate the intents and reasons that drive users' query reformulation behavior. We also analyze how users interact with different types of query reformulation entries as well as how systems should better support users in reformulating queries. Specifically, we aim to address the following research questions:

- **RQ1**: How do users' query reformulating behaviors evolve within search sessions?
- **RQ2**: Do users perform reformulations differently under various search intents?
- **RQ3**: Can we predict why and how users reformulate their queries?

To shed light on the above research questions, we conducted a large-scale field study to collect both implicit user behavioral signals as well as the first-tier feedback concerning query reformulation. An in-depth investigation of users' evolving reformulations within sessions is then presented. We further compare user reformulating behaviors across various search intents and figure out subtle differences between user actions. To take further steps in modeling user reformulations, we propose two novel challenges of predicting why and how users reformulate a query. Inspired by the findings in our field study, we design a supervised learning method. Extensive experimental results on our collected data have shown that our method can achieve high-quality prediction in both tasks.

In summary, the contributions of this work are three folds:

- We conduct a field study tailored for the investigation of user reformulations and obtain a large-scale practical dataset with abundant daily search behaviors. The dataset supports more in-depth investigations of user reformulating behaviors [1].
- We thoroughly analyze the trends of various aspects of users' reformulating behaviors and the corresponding distributions under different intent taxonomies. The findings provide new insight into complex user reformulations as well as the guidelines for designing more efficient query suggestion techniques in SERPs.
- Based on the analysis in the field study, we propose a supervised learning framework to predict why and how users reformulate a query, both of which are novel challenges in this domain. Extensive experimental results have shown the possibility of modeling more delicate aspects of user reformulations by leveraging session contexts.

[1]Now available at http://www.thuir.cn/tiangong-qref

2 RELATED WORK

2.1 Query Suggestion and Auto-completion

Whether users make good use of search engines depends largely on whether they submit appropriate queries. To assist users in reformulating their queries, numerous studies have aimed at improving services like query suggestions and auto-completion [1, 5, 11, 29, 35]. Early work relied on the query association or similarities within search sessions, e.g., mining the association rules [10, 20, 36] or the co-occurrence relationships [15]. Based on Markov Models, Cao et al. [1] learned the associations between consecutive queries and proposed a novel query suggestion model, namely QVMM. Another body of research leverages statistical features to learn user reformulating behaviors [6, 32, 34]. By analyzing the trends of user reformulations within sessions, Jiang et al. [21] extracted features including the co-occurrence frequency, position in the session, and so on, to improve query auto-completion. With the boom of deep learning, Sordoni et al. [35] first adopted the hierarchical RNN-based framework to encode the query history within sessions. Dehghani et al. [7] incorporated a copy mechanism into the query decoder of their model given the observation that most query terms within a session are retained from the previously submitted queries.

While these frameworks are effective in predicting the next queries that users may issue, they could not explain how a contextual factor will spur on a reformulating action, i.e., lack of interpretability. More in-depth investigations should be conducted to better understand why and how users reformulate a query.

2.2 User Reformulation Analysis

To better understand user reformulations, there is a broad spectrum of research aiming at analyzing users reformulating behaviors in various search scenarios [2, 14, 17, 25, 31]. Huang et al. [16] investigated various reformulation strategies of search users based on a search engine log. These strategies are based on the content change, including word reorder, remove/add words, URL striping, acronym, substring/superstring, abbreviation, etc. To enhance E-commerce search, Hirsch et al. [14] analyzed the distribution of different types of reformulations, changes of search result pages retrieved for the reformulations, and clicks&purchases performed upon the retrieved results on the eBay platform. From other perspectives, reformulating behaviors have also been leveraged to predict satisfaction [12, 18] as well as to model session contexts [28].

Web search involves abundant human interactions, hence research in Information Retrieval (IR) should properly consider user perceptions. Besides analyzing search logs, user studies have also been conducted to explore the way that users formulate their queries [9, 23]. Based on eye-tracking, Eickhoff et al. [9] studied query refinement to gain precise and detailed insight into which terms the user was exposed to in a search session. Compared to previous ones, their work provided more insight into finer-grained user reformulations. However, they focused merely on the content of query reformulations, i.e, tracking user attention at the term level. Also, experimental setups in their lab-based user study may

cause participants to perform differently from the realistic scenarios. To take further steps in understanding user reformulation, we conducted a long-term field study to collect a practical search dataset that is available for more in-depth investigations on user reformulations.

3 FIELD STUDY

Field study has been proposed to overcome the limitations of lab studies and large-scale log analysis [13]. Experimental setups in user studies in controlling some factors may cause participants to behave differently while search engine logs are usually noisy and only contain observations (e.g., queries, timestamps, clicked results, and so forth) from users. Therefore, we conducted a one-month field study to collect more realistic and detailed behavioral data as well as first-tier user feedbacks. Our study is inspired by previous work [37, 38], but focuses more on detailed facets of user reformulations. Zhang et al. [38] conducted a study to explore the consistency between user behavior modeling and satisfaction measurement to better understand evaluation metrics, while Wu et al. [37] designed the study specifically for the image search scenario. To emphasize on detailed user reformulating behaviors and their cognitive process in session search, we collected more user feedback (e.g., why and how they reformulate a query) as well as implicit actions, which were not considered in previous work.

3.1 Procedure

Following previous work [13, 37], the procedure of our study mainly consists of three stages:

3.1.1 Task Introduction. We recruited 50 participants via an online pre-experiment questionnaire which collected their demographic information and daily search habits. All of them were familiar with the basic usage of search engines. After signing a consent form and agreeing with our data collection policies, participants applied for our field study and were instructed with the requirements via an online meeting. They were also told to install a browser extension on their PCs (desktops or laptops). The extension was tailored for our study purpose, which recorded the participants' daily search activities and could be turned on or off at any time. After the task introduction, we further invited them to complete a 10-minute pilot study. Having ensured that all participants were familiar with the experimental process and understood some key concepts, they were told to use their PCs for daily search anywhere as usual.

3.1.2 Data Collection. Our lab study lasted for about one month. During this period, the participants' daily Web search activities would be recorded automatically by the browser extension if it was on. Participants could review previously issued queries, split them into search tasks, and further provide feedback at their convenience. To ensure the quality of user annotations, we set a two-day expiration for all search tasks and queries (we suppose that a participant may not have distinct impressions on a task after this time interval). If a search task had not been reviewed within two days after it was recorded, then all queries within it together with their corresponding search logs would be removed from our database. To ensure individual privacy, the participants could remove any recorded queries at the reviewing phase. More details of the collected data will be later introduced in Section §3.2.

3.1.3 Summarization. After one month of data collection, we announced the completion of the field study to participants. They were remunerated according to their engagement and contribution: about $6 for the basic participation plus $0.15 for each valid query log. According to a simple post-experiment interview, most participants were satisfied with the design of our field study as well as the remuneration, indicating the rationality of the experimental setups.

3.2 Experimental Platform & Data Description

Our experimental platform consists of: 1) a Chrome extension [2], and 2) an annotation platform, of which the former records the participants' daily search activities while the latter collects their feedbacks. An overview of explicit information we collected is presented in Table 1.

3.2.1 Search Behavior Log. The Chrome extension we developed could be installed on various chrome-based browsers and could record search-related activities when specific events such as clicking or mouse movements were triggered. To better understand how users reformulate a query, the extension recorded the sources of reformulations by locating the action within the current SERP. Other information we recorded are listed as follows: 1) **HTML**: including the URLs and HTML contents of SERPs and landing pages; 2) **Mouse events**: including details about mouse movements, clicks, and scrolling; 3) **Queries**: the content of queries that the participants issued; 4) **Timestamps**: including the starting and ending timestamps for all pages and user activities.

3.2.2 Search Feedback. As the browser extension recorded user activities implicitly, we also developed an annotation platform to collect more explicit feedback from users. Our annotation platform mainly consists of five functional screens (each screen collected some information, as shown in Table 1). While reviewing the search task, the participants needed to go through these screens sequentially, yet they could leave the pages at any time and then continue annotating by re-entering from the home page. The five screens are as follows:

- **I: Search task identification**: In this screen, participants needed to review the queries they have issued and group them into sessions according to search intents. Compared to previous work that adopted a time threshold (i.e., 30 minutes) to split sessions [3, 21], this approach may be more appropriate. Moreover, they can freely remove a query no matter whether it has been assigned to a task. This would make them feel relaxed and act as usual while searching.
- **II: Task annotation A**: In this screen, participants needed to fill a post-questionnaire about the search task, which includes the annotation about: 1) urgency, 2) atmosphere, 3) specificity, 4) trigger, and 5) expertise. The details of these attributes are presented in the top five rows of Table 1.
- **III: Query annotation**: For a specific search task, participants should provide feedbacks for each query. More descriptions about each question can be found in the "Query" row group in Table 1. Based on the previous work [16, 31], we create a new intent-level taxonomy for reformulation types. As existing taxonomies may be coarse-grained and some types are overlapped with the others

[2] Support for two largest commercial search engines in China: *Baidu* and *Sogou*.

Table 1: Descriptions of explicit information collected in our field study. The superscript $1/2$ denotes the attribute was collected via the annotation platform/browser extension. Implicit signals were also collected but are not included in this table.

	Attribute	Description	Screen	Value / Option
Task	Urgency[1]	Were you very urgent in completing this search task?	II	(0) have plenty of time - (4) very urgent
	Atmosphere[1]	How was the environment while you were searching?	II	(0) very silent - (4) very noisy
	Specificity[1]	How specific was your search intent?	II	(0) very broad - (4) very clear
	Trigger[1]	†How was this search task motivated?	II	(0) interest-driven - (4) task-driven
	Expertise[1]	Were you familiar with the search tasks before searching?	II	(0) not at all - (4) very familiar
	Satisfaction[1]	Were you satisfied with the search process during this task?	V	(0) unsatisfied - (4) very satisfied
	Difficulty[1]	How do you feel about the difficulty of finding useful information?	V	(0) very easy - (4) very difficult
	Success[1]	Did you find any useful information for this task?	V	(0) almost no - (4) all you want
Query	Reformulation Type[1]	What is the intent-level relation between this query and the last one?	III	(A) specification; (B) generalization; (C) meronym; (D) holonym; (E) synonym; (F) somewhat relevant¶; (G) total new topic; (H) other:__
	Reformulation Reason[1]	Why did you reformulate this query or end your search?	III	(A) having found enough information; (B) found no useful information with great effort; (C) come up with a better query, w/o intent shift; (D) come up with a more interesting query, w/ intent shift; (E) other:__
	Reformulation Entry[2]	The interface that the user used to reformulate this query.	–	(A) search input box; (B) query suggestion (related queries); (C) related entities; (D) hot or top searched queries; (E) entries in other pages
	Reformulation Inspiration[1]	Which component inspired you to reformulate this query?	III	(A) SERP search snippets; (B) SERP other components; (C) landing pages; (D) others (e.g., a cognitive snap)
	Satisfaction[1]	Were you satisfied with the search results in this query?	III	(0) unsatisfied - (4) very satisfied
Result	Usefulness[1]	How do you rate the usefulness for each result for completing the search task?	IV	(0) useless - (2) highly useful; (3) serendipity

† This attribute describes the weights of two motivations: search for interest or task completion. Lower values mean a user is mainly motivated by interest while higher values mean the search is mainly task-driven. A medium value represents the search process is both interest- and task-driven. We suppose that a search process with neither interest nor task purpose makes no sense.
¶ Participants were told to select this option if they found a query was related to the last one but the relationship did not belong to option (A) - (E);

at the intent level, we made two changes here: 1) we merge various types with the same search intents into "synonym", e.g., same queries, spelling correction, reorganizing query terms, abbreviation, etc; 2) some types are divided into finer-grained subtypes, e.g., "specification"→"specification" and "meronym". For example, "iPhone X" is a specialized query of "iPhone" while "iPhone screen " is the meronym of it. However, many previous studies roughly classified them into "specification" without distinction. We deem that these two relationships are different and the new taxonomy can better distinguish subtle differences between users' evolving intents. To better understand users' evolving intents, we also collected the reason, inspiration source for each reformulating behavior. Detailed descriptions for some options in these attributes are presented in Table 2. At the end of each query annotation component, there is an entry for the next screen: SERP annotation. Having annotated all queries and clicked the submission button, the screen will locate at V.

- **IV: SERP annotation**: This screen is a child window of each query in screen III, providing the view of the corresponding SERP. Participants should refer to the view and rate the usefulness for each result using a 4-point Likert-type scale from 0 to 3. Results with the highest usefulness are denoted as "serendipity", which represents that not only do they contain useful information but also bring surprise to participants. To alleviate the participants' efforts, participants were told only to rate usefulness for the results they had examined. After submitting the annotation, the screen will return to III.
- **V: Task annotation B**: Including the annotation for task satisfaction, difficulty, and success. Having submitted this page, all information for the search task will be saved in our database.

Table 2: Detailed explanations for some options in reformulation type, reason, entry, and inspiration source.

Attribute	Option	Example/Distinctions Between Other Options
Type	(A)	iPhone→ iPhone X; desktop wallpaper→ desktop wallpaper HD
Type	(B)	iPhone X→ iPhone; desktop wallpaper HD→ desktop wallpaper
Type	(C)	iPhone→ iPhone battery
Type	(D)	iPhone battery→ iPhone
Type	(E)	abbr→ abbreviation; laebl→ label; capital of Slovenia→ Slovenia capital
Reason	(A)	Users left this query with **satisfaction**.
Reason	(B)	Being **unsatisfied** with most results, users **were forced to** change the query.
Reason	(C)	Users **initiatively** came up with a better query to fulfill **the current search intent**.
Reason	(D)	Users' intent **shifted** to other subtopics/topics.
Insp.	(A)	Including the contents of titles and snippets.
Insp.	(B)	Including other components in SERP except for search snippets, e.g., query suggestion.
Insp.	(D)	The inspiration was not from the screen.

Participants could continue entering screen I to annotate other tasks or leave the platform.

3.3 Participants and Collected Data

We filtered a proportion of data since some information was not recorded accurately. Moreover, we found invalid annotations from two participants through manual inspection. There were also five participants having searched nothing. After data cleaning, we reserved information from 43 participants. These participants were aged from 18 to 52, of which 22 were male while the rest were

female. Among them, there were 17 undergraduates, 16 graduates, and ten employees from different universities and companies.

Finally, we collected 5,958 sessions and 12,752 queries in total. The number of sessions that contain more than one query is 2,356, of which short sessions (length=2) account for about 46.7%. The distribution of session length is presented in Figure 2. Compared to search logs [3, 30], our data contains more longer sessions. As we focus on user reformulating behaviors, only sessions that contain at least two queries were considered for our study. Overall, each participant clicked 1.04 results and acted 0.29 clicks on other components in the SERP per query. Meanwhile, they browsed 2.59 landing pages or other pages on average for each query.

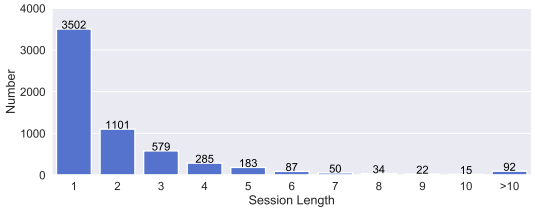


Figure 2: Distribution of session length in our collected data.

4 USER REFORMULATION ANALYSIS

Based on the collected data, we delve into two main aspects to better understand user reformulating behaviors. In Section §4.1, we first answer **RQ1** by analyzing the overall trends of user reformulations within sessions. As long sessions may contain much noise, we only consider sessions with no more than ten queries, which account for over 95% of all sessions. We further make a comparative analysis of fine-grained reformulating behaviors under different search intents in Section §4.2 to investigate **RQ2**.

4.1 Reformulation Evolution Within Sessions

In this section, we analyze how various aspects of user reformulation evolve over time in search sessions. These aspects include: 1) reformulation type at both syntactic- and intent-level, 2) reason for reformulation, 3) reformulation entry, and 4) inspiration source of reformulation (See descriptions of the aspects in Table 1).

4.1.1 Reformulation type. Based on our new intent-level taxonomy, we have collected information on reformulation type through the field study. To compare the difference between our taxonomy with the previous ones, following [3, 16], we also develop a syntactic taxonomy by analyzing the query change. Let q_t/q_{t-1} be two consecutive queries and $W(q)$ be the bag of words for q, we define added/deleted/intersected terms as follows:

$$+\triangle q_t = \{w|w \in W(q_t), w \notin W(q_{t-1})\};$$
$$-\triangle q_t = \{w|w \notin W(q_t), w \in W(q_{t-1})\};$$
$$\cap q_t = \{w|w \in W(q_t), w \in W(q_{t-1})\}.$$

Then the five types can be formulated as:

- *Add*: $+\triangle q_t \neq \emptyset, -\triangle q_t = \emptyset$;
- *Delete*: $+\triangle q_t = \emptyset, -\triangle q_t \neq \emptyset$;
- *Change*: $+\triangle q_t \neq \emptyset, -\triangle q_t \neq \emptyset, \cap q_t \neq \emptyset$;
- *Repeat*: $+\triangle q_t = \emptyset, -\triangle q_t = \emptyset, \cap q_t \neq \emptyset$;
- *Others*: $+\triangle q_t \neq \emptyset, -\triangle q_t \neq \emptyset, \cap q_t = \emptyset$.

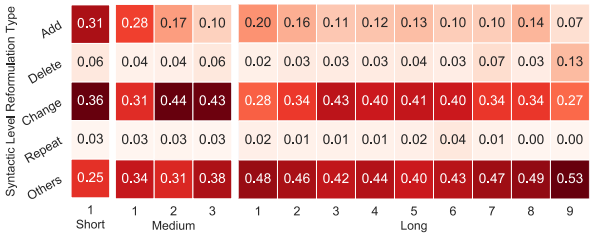


Figure 3: Distribution of reformulation type at syntactic level across reformulation steps in three kinds of sessions.

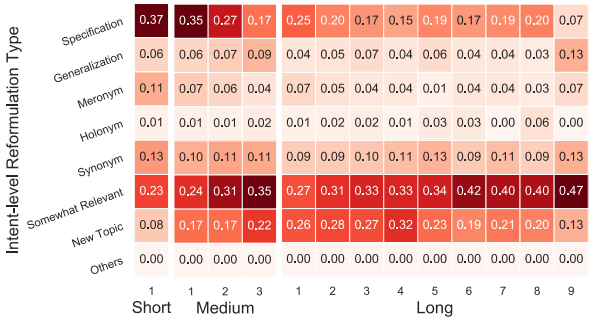


Figure 4: Distribution of intent-level reformulation type across reformulation steps in three kinds of sessions

We divide all sessions into three groups according to their lengths: 1) short sessions (with two queries); 2) medium sessions (with three to four queries); 3) long sessions (with at least five queries). The evolving distribution of user reformulation type at the syntactic level is presented in Figure 3.

From Figure 3, the most frequent reformulating modes are "*Change*", "*Add*", and "*Others*". We also find a stable decay of the "*Add*" type from the first to last reformulation step in all kinds of sessions, indicating that users tend to narrow down the search scope by adding constraints on queries at the beginning of a session. As the search progresses, users become gradually clear about their search intents and are less likely to continue adding constraints. They gradually shift their intents to other subtopics by replacing some terms from the previous queries (a rise of "*Change*" action from 0.31 to 0.44 in medium sessions, from 0.28 to 0.43 in long sessions) or even new topics by issuing very different queries (a great proportion of the "*Others*" type in the later steps). Besides, compared to short and medium sessions, there is a smaller proportion of "*Add*" actions in long sessions, which may be due to the ambiguity of intent in complex tasks. The difficulty of reformulation can be the direct cause for long sessions and should be emphasized.

As the syntactic taxonomy is coarse-grained, we also plot the distribution of our intent-level taxonomy in Figure 4. Compared to Figure 3, we find that the distribution is more balanced. Furthermore, there is almost no "*Others*" type annotated by users across reformulation steps. These observations imply that our taxonomy can better cover various conditions of user intent change while reformulating. After merging several types with the same search intent into "*Synonym*" (See Table 1), we find that at all steps the

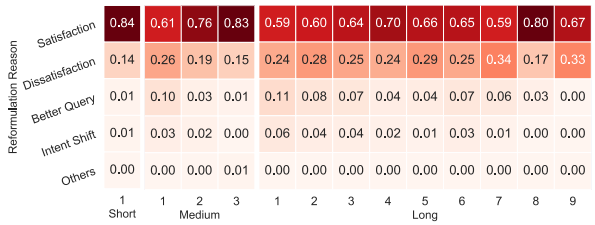


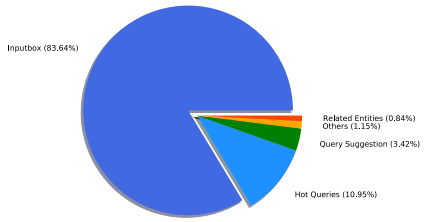
Figure 5: Distribution of reformulation reason across reformulation steps in three kinds of sessions

probability that users keep their search intents is around 10%. Intuitively, we expect the trend of "*Somewhat Relevant*" will be similar to "*Change*" in the syntactic taxonomy. However, we find a difference that the proportion of "*Somewhat Relevant*" steadily increases across reformulation steps in all session conditions. Other subtle differences between similar types of two taxonomies can also be found, such as "*Add*" (average 0.29) vs. "*Specification*" (average 0.36), and "*Others*" (↘↗) vs. "*New Topic*" (→↗↘), especially in long sessions. Nevertheless, some trends are similar, such as "*Add*" (↘) vs. "*Specification/Meronym*" (↘), and *Delete*" (↗) vs. "*Generalization/Holonym*" (↗).

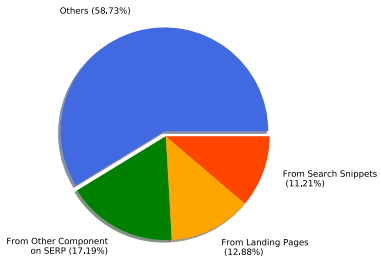
Generally, according to the tendencies in two figures, we can summarize Web search into a two-phase process: *Specialization* → *Intent Shift*. In the first phase, users specialize their queries to focus on a more detailed facet of the search purpose. In the second phase, they shift their intents to other subtopics or a new topic.

4.1.2 Reformulation reason. Understanding why users reformulate their current queries may guide to design better query suggestion techniques on SERPs. According to the observations in the pilot study, we defined four reasons for a reformulation, which are described as options from (A) - (D) in Table 2. To better plot the trend of the reasons why people reformulate the current query, we label options (A) - (D) as "*Satisfaction*", "*Dissatisfaction*", "*Better Query*" and "*Intent shift*" according to user intent and satisfaction. As shown in Figure 5, we observe that most users leave the current query because they have found enough useful information. With more queries being issued, they are more likely to be satisfied with search results (Satisfaction: "↗"; Dissatisfaction: "↘"). This finding is consistent with our analysis of the reformulation type. Users become increasingly clear about the task during search and could formulate more appropriate queries that retrieve satisfying results for them. We also find downtrends for "*Better Query*" and "*Intent Shift*" across session iterations, indicating that users prefer to replace query terms at the very beginning of sessions. Being more satisfied after several search rounds, users may not need to strive for better queries anymore.

4.1.3 Reformulation entry. Typical reformulation entries provided by modern search engines include: search input box, query suggestion, related entities, hot queries, and so on. As queries within each entry component represent a specific direction of user intent change, we should also analyze the entries that users adopt for reformulation. Figure 6(a) plots the overall proportion of each reformulation entry that users adopt in daily search. According to the pie chart, users submitted most queries via the search input box (accounting for 83.64% of all reformulating behaviors), followed by hot queries (10.95%). To our surprise, the proportions of query suggestions and related entities are only 3.42% and 0.84%, respectively. This phenomenon shows that search users seldom use the interfaces provided by search engines for reformulation. The frequency of entering a new query from related entities is even lower than from other pages. We further investigate the overall trend of the utilization of several functional reformulation entries across session reformulation steps. The left subfigure of Figure 7 shows a decline of users' clicking into the query suggestions over the session iterations. On the contrary, they are more likely to be attracted by hot queries at any step.



(a) Proportion of reformulation entries that users adopt.



(b) Proportion of the reformulation inspiration.

Figure 6: Proportion of reformulation entries adopted by users and the inspiration source for each reformulation.

4.1.4 Inspiration source. To save users' effort, search systems should better guide them in query reformulation. Before that, it is of vital importance to understand how their reformulating behaviors can be influenced by search engines. While browsing Web pages for the current query, some factors on the pages may inspire a user to formulate a new one. Therefore, we aim to investigate the inspiration source of each reformulating behavior. Different from the results in Section §4.1.3, there are about 17.2% reformulations influenced by other components on SERP in Figure 6(b). This gap suggests that although these entries are not frequently accessed by users, they provide some inspirations for them. Another interesting finding is that users are likely to be inspired by search snippets and landing pages equally, which implies the effectiveness of search snippets in enlightening users and the convenience they provide so that users are unnecessarily to click into landing pages. However, the majority of the reformulation inspirations (about 58.7%) come from neither SERPs nor landing pages. As shown in the right subfigure in Figure 7, user effort in reformulation is huge across all steps, especially in long sessions (e.g., 87% of "*Others*" inspiration source in the end). Users are likely to depend more on themselves when reformulating a query in longer sessions. There may be two main

reasons: 1) search engines do not provide sufficient guidance for users in complex or in-domain topics; 2) the quality of existing query suggestions cannot fully satisfy users' search purposes.

4.2 Reformulation With Various Intents

User intent and domain expertise will influence their search process. For instance, a broad intent may cause a search process with more exploration. A clear intent, on the contrary, can lead to more exploitation. To investigate how users' search intents would translate to their reformulating behaviors, we make a detailed comparison on the search effort & gain, the proportion of each reformulating actions, and the user actions under various intents. The Dimensions we consider for search intents include the trigger and specificity, which have been collected through our field study. As domain expertise has also been proved to impact user interactions and search outcomes [27, 39], we also analyze the query reformulation behaviors of users with different expertise levels. Regarding the dependencies between intent trigger, intent specificity, and domain expertise, we only find a low positive correlation between trigger and expertise (Kendall's $\tau = 0.3$, $p < 0.001$), indicating that if a session is triggered by some task to a greater extent the user may be more familiar about the search domain. Other dimensions can be considered to be independent of each other. For the sake of analysis, we introduce the taxonomies of three dimensions with their corresponding values as follows:

- **Intent Trigger**: Interest-driven(0), Interest- and Task-driven (2), Task-driven (4);
- **Intent Specificity**: Clear (3-4), Borad (0-1);
- **Domain Expertise**: Familiar (3-4), Unfamiliar (0-1).

Following [37], here we ignore ambiguous values (e.g., the median values) to better categorize intents. The differences in user search effort and gain, reformulating behaviors, and user actions with regard to various intent triggers, specificity, and expertise are presented in Table 3. As our data is not normally distributed, we use the Kruskal-Wallis test [26] to calculate the significance. We further calibrate each p-value with the Bonferroni Correction [33] to control the Family-Wise Error Rate.

Generally, through the comparison between different dimensions, we find that all dimensions have a great impact on users' search effort and gain, especially in the query-related behaviors. However, there are few differences in reformulation across different levels of domain expertise, as we only observe significances in four variables, e.g., "*%Type=Specification*" and "*%Inspiration=Others*". This suggests that users' familiarity with the search domain is less prone to affect their reformulation patterns. On the other hand, the intent trigger has the largest impact on reformulating behaviors as well as user actions, followed by the specificity of intent. In the third column of Table 3, we can observe that most variables vary significantly across different intent triggers. Therefore, we will focus more on this dimension later.

4.2.1 Intent trigger. In interest-driven tasks, users take less effort in the search process (e.g., fewer task queries and shorter task time). Task difficulty perceived by users is lower (0.86 vs. 1.34 in task-driven sessions) while the average satisfaction and success are both higher than the other two taxonomies. Moreover, they tend to submit queries with a high unique term ratio (0.93) under diverse intents ("*%Type=New Topic*": 0.19). It seems that users depend more on search engines for reformulation, as they access hot queries and query suggestions more frequently, making the submitted queries more diverse. They are also more likely to be inspired by the contents on SERPs to organize the next queries (other components on SERP: 0.23, search snippets: 0.19). On the contrary, task-driven sessions are significantly more difficult. It is still hard for users to achieve search success (3.14 vs. 3.48 in interest-driven tasks) even when they have paid greater efforts. As users need to pay more effort in reformulation ("*%Inspiration=Others*": 0.59), perceived task satisfaction is relatively lower on average ("*task satisfaction*": 3.16,). The unique term ratio is relatively low, indicating there are more overlapped terms between consecutive queries. This may be because users tend to submit queries that are more relevant to the previous ones to better find useful information for task completion. To do so, they are always searching for a better formulation of their information need ("*%Reason=Better Query*": 0.11) by attempting more generalized queries, synonyms, and intrinsically diverse queries.

For the sessions that are both interest- and task-driven, most values are the interpolation between two extremes. Nevertheless, we find highest values in 1) *% Type=Specification*, 2) *% Interface=Search Box*, and 3) *% Reason=Satisfaction*. In this condition (a typical scenario can be online shopping), people are more willing to narrow down the search scope step by step until they have found enough useful information. They are also found to browse the Web pages more carefully, i.e., with slower mouse movements. Both search interest and the purpose of task completion may motivate users to be more engaged in the search process. Finally, although the tasks are more difficult than interest-driven only sessions, users can still achieve a similar level of success. In summary, users are mostly guided by search engines in sessions that are only triggered by interests, followed by sessions that are both interest- and task-driven. In long and complex tasks, we observe relatively weak guidance from search engines in assisting users to better reformulate queries.

4.2.2 Intent specificity. Users with broad search intents tend to submit shorter queries while spending more time on all pages. With an ambiguous intent, they may need to pay more attention to the contents in SERPS or landing page to extract useful information. In comparison, tasks started by clear search intent are more difficult on average. However, these users achieve comparable search success and tend to be more satisfied when leaving the current query. This may be because users with clear intentions have a better understanding of the task nature thus can judge the difficulty of the task more accurately. Even if the difficulty is greater, the search task can be successfully completed by them. These users are also significantly more likely to be inspired by search engines, suggesting that they can better utilize the services in search engines.

4.2.3 Domain expertise. Users who are familiar with the search task pay more effort on average due to the high task difficulty, while it is easier for unfamiliar users to reach their search goal. However, domain expertise does not have significant effects on user reformulations. One phenomenon is that in-domain users formulate more specialized queries with less inspiration from search engines. It indicates that users who are familiar with the task are more inclined to rely on their own knowledge for query reconstruction and gradually narrow down the search scope.

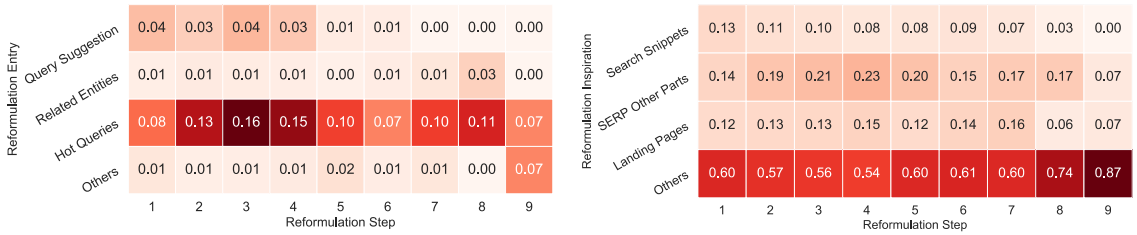


Figure 7: Overall distributions for the user utilization of different entries (left, the row of search box is omitted to better focus on other functional entries) and the inspiration source for reformulation (right) across session reformulation steps.

Table 3: Differences in user search effort & gain, reformulating behavior, and user actions w.r.t. different intent triggers, specificity, and expertise. "*//***" indicates a statistical significance at $p < 0.05/0.01/0.001$ level using Kruskal-Wallis's H Test among different taxonomies in one dimension. Note that all p-values are calibrated through Bonferroni correction within the corresponding behavior group. Underlines in the "Intent Trigger" dimension denote a non-significant difference ($p > 0.05$) between the values in two taxonomies using Dunn's post-hoc test [8] although the difference within the dimension is significant.**

	Behavioral Variables	Intent Trigger				Intent Specificity			Domain Expertise		
		Interest-driven (0)	Interest- & Task-driven (2)	Task-driven (4)	Sig.	Clear (3-4)	Broad (0-1)	Sig.	Familiar (3-4)	Unfamiliar (0-1)	Sig.
Search Effort & Gain	*# task queries*	1.68	2.24	**2.77**	***	2.10	2.36	–	**2.53**	1.87	***
	# task query terms	7.32	10.0	**13.1**	***	**11.5**	9.2	***	6.52	**10.9**	***
	# task unique terms	6.30	6.94	**8.64**	**	**10.2**	6.52	***	**8.72**	7.94	***
	# avg. terms per query	4.25	**4.52**	4.16	***	**4.65**	4.17	***	**4.55**	3.77	***
	# avg. unique terms per query	**3.90**	3.66	3.16	***	**4.37**	3.46	***	**3.97**	3.13	***
	unique term ratio	**0.93**	0.84	0.83	***	**0.94**	0.86	***	0.89	0.88	–
	task time (s)	404.1	415.7	461.0	–	311.8	**442.2**	***	**429.0**	404.3	*
	avg. dwell time on SERP (s)	175.9	143.3	164.0	–	130.4	**167.1**	***	**174.8**	149.7	***
	avg. dwell time on other page (s)	70.63	112.9	136.5	–	60.74	**106.7**	***	90.3	107.0	–
	# total pages	6.28	5.76	5.76	–	5.72	6.04	–	5.84	6.05	–
	# clicks on SERPs	**4.47**	3.99	2.78	***	4.30	3.62	–	**4.07**	3.46	***
	# clicks on other pages	0.36	0.39	0.73	–	0.22	0.54	–	0.39	**0.59**	***
	# clicks on search results	3.38	**3.39**	2.06	***	2.74	**2.99**	–	**3.33**	2.31	***
	# clicks on others	**1.45**	1.00	1.45	***	**1.77**	1.16	**	1.14	1.75	–
	task satisfaction (0-4)	**3.48**	3.34	3.16	***	3.40	3.37	–	3.32	**3.46**	***
	task difficulty (0-4)	0.86	1.28	**1.34**	***	**1.71**	1.01	***	**1.29**	0.91	***
	task success (0-4)	**3.48**	3.34	3.14	***	3.36	3.36	–	3.30	**3.46**	***
Reformulating Behaviors	*% Type=Specification*	0.27	**0.40**	0.22	***	0.22	**0.30**	**	**0.33**	0.24	***
	% Type=Generalization	0.03	0.04	**0.06**	*	0.03	**0.06**	*	0.07	0.05	–
	% Type=Synonym	0.09	0.11	**0.17**	***	0.05	**0.13**	***	0.11	0.13	–
	% Type=Somewhat Relevant	0.19	0.24	**0.31**	***	0.19	0.24	–	0.19	**0.25**	*
	% Type=New Topic	**0.19**	0.06	0.12	***	**0.30**	0.10	***	0.12	0.16	–
	% Entry=Search Box	0.76	**0.94**	0.91	***	0.70	**0.90**	***	0.85	0.84	–
	% Entry=Hot Queries	**0.15**	0.02	0.05	***	**0.24**	0.05	***	0.09	0.09	–
	% Entry=Query Suggestion	**0.07**	0.03	0.03	***	0.05	0.04	–	0.05	0.04	–
	% Insp=Others	0.30	0.58	**0.59**	***	0.28	**0.57**	***	**0.54**	0.45	**
	% Insp=SERP Other Components	**0.23**	0.06	0.10	***	**0.37**	0.09	***	0.12	**0.17**	*
	% Insp=Landing Pages	0.11	0.14	0.09	–	0.07	0.12	–	0.12	0.09	–
	% Insp=Search Snippets	**0.19**	0.11	0.11	***	**0.17**	0.11	*	0.12	0.14	–
	% Reason=Satisfaction (A)	0.66	**0.71**	0.55	***	**0.80**	0.66	***	0.69	0.68	–
	% Reason=Dissatisfaction (B)	0.21	0.20	**0.30**	***	0.08	**0.23**	***	0.21	0.21	–
	% Reason=Better Query (C)	0.09	0.07	**0.11**	*	0.08	0.08	–	0.07	0.08	–
User Actions	*# avg. clicks per query*	**1.64**	1.46	1.19	***	1.44	1.44	–	**1.51**	1.39	***
	avg. click rank	2.51	2.43	2.41	–	2.31	2.53	–	2.46	2.48	–
	max. click rank	4.06	4.08	3.75	–	3.75	4.16	–	4.04	3.92	–
	avg. annotated usefulness (0-3)	0.25	**0.26**	0.20	***	0.20	**0.25**	***	**0.26**	0.20	***
	moving time ratio	0.45	0.46	0.42	–	0.43	0.46	–	0.41	**0.47**	*
	scrolling time ratio	**0.28**	0.27	0.21	***	0.23	**0.26**	–	0.24	**0.25**	–
	avg. moving distance (pix)	59.9	52.4	**61.5**	***	58.3	57.5	–	57.2	57.1	–
	avg. moving speed (pix/s)	**561.7**	507.9	560.0	*	554.3	537.4	–	566.6	521.0	–
	avg. scrolling distance (pix)	81.9	63.9	**84.0**	***	**86.2**	72.6	***	68.7	**83.4**	***
	avg. scrolling speed (pix/s)	**921.2**	699.0	860.9	***	**943.9**	787.7	***	757.6	**891.0**	***

4.3 Summary

In this section, we have analyzed the trend of reformulation revolution within sessions as well as the effects of search intents in different taxonomies on user reformulating behaviors to address **RQ1** and **RQ2**, respectively.

To answer **RQ1**, we conclude as follows: 1) User search process can be summarized into a two-phase process: *Specialization→Intent*

Shift. Users tend to add more constraints on existing queries at initial steps to narrow the scope of the following search. As their search intents are gradually satisfied, they will start to shift their intents intrinsically or extrinsically. 2) Users mainly use the search box for query reformulation. Whereas, a certain proportion (about 40%) of their inspirations are from SERPs and landing pages, indicating that there is potential for search engines to better provide query suggestion services by leveraging session contexts. 3) As users pay too much effort on complex search tasks, search engines should be improved in guiding users to better reformulate their queries. Our results highlight the necessity of search engines to leverage search contexts for query recommendation.

To answer **RQ2**, we conclude that: 1) Users' reformulating behaviors can be largely influenced by both the trigger and the specificity of their intents, while their domain expertise has less effect. 2) Existing query suggestion services in SERPs benefit interest-driven tasks more. However, in task-driven sessions where users need to pay more efforts, search engines only provide limited assistance. Through user modeling, search engines may better identify the intents behind user behavioral signals and further provide personalized query suggestion services for various intents.

5 PREDICTING VARIOUS ASPECTS OF REFORMULATING BEHAVIOR

Better understanding and further predicting users' reformulating behaviors are beneficial for the optimization of search systems. Most previous work only consider reformulation contents, i.e., predicting the next queries that users might issue. However, it may provide little help for alleviating users' search efforts by merely fitting the user query sequences. We need to explore detailed aspects such as why and how they would formulate queries. To this end, we address two novel challenges in **RQ3** by predicting: 1) why users leave the current SERP and reformulate the query, and 2) how they reformulate the current query. To better tackle the second challenge, we split it into two subtasks: a) whether users will use other reformulating entries except for the search box; and b) if so, which entry they will access. Finally, we obtained three subtasks, among which two were four-class classification problems and the other one was a binary classification problem. For simplicity, we denote the three tasks as: *Why*, *Whether*, and *Which* problems in what follows.

5.1 Features

In this work, we propose a supervised learning approach for prediction. Based on the analyses in Section §4, we find that various search intents have a great impact on users' reformulations as well as other behavioral variables such as query, dwell time, click, and their mouse activities. We prefer that these interrelated variables can contribute to the prediction of user reformulations. Therefore, we extract several features for each group based on the observations within sessions since annotations can not be directly obtained in realistic scenarios. Previous studies have also highlighted the effectiveness of using session-level trends for query auto-completion [21]. Therefore, we also introduce several tendency-based features. All features with the corresponding descriptions are given in Table 4.

Table 4: Descriptions of all features that we use.

Group	Feature
Query	Q1 - number of previous queries
	Q2 - proportion of unique queries
	Q3 - number of terms in the current query
	Q4 - average number of terms in previous queries
	Q5 - average number of unique terms in previous queries
	Q6 - average Jaccard similarity of previous queries
	Q7 - average Levenshtein distance of previous queries
Dwell Time	D1 - dwell time on the current query
	D2 - dwell time on previous queries
	D3 - total dwell time on SERPs in previous queries
	D4 - average dwell time on SERPs in previous queries
	D5 - total dwell time on other pages in previous queries
	D6 - average dwell time on other pages in previous queries
Click	C1 - number of clicked results in the current SERP
	C2 - number of clicks on other components in the current SERP
	C3 - number of clicked results in previous SERPs
	C4 - number of clicks on other components in previous SERPs
	C5 - average number of clicked results in previous SERPs
	C6/C7 - min/max clicked rank in the current query
	C8/C9 - min/max clicked rank in previous queries
Mouse	M1/M2 - average moving distance/speed in the current query
	M3/M4 - average moving distance/speed in previous queries
	M5/M6 - average scrolling distance/speed in the current query
	M7/M8 - average scrolling distance/speed in previous queries
	M9 - max browse depth in the current SERP
	M10 - max browse depth in previous SERPs
Trend	T1/T2 - trend of Jaccard similarities in previous queries
	T3/T4 - trend of Levenshtein distances in previous queries
	T5 - Trend of query dwell time in previous queries

Let q_T be the current query and t_T be the dwell time on q_T, then the features in the "Trend" group can be formulated as:

$$T1 = Jaccard(q_{T-1}, q_T) / \frac{1}{T-1} \sum_{i=2}^{T} Jaccard(q_{i-1}, q_i);$$

$$T2 = Jaccard(q_{T-1}, q_T) / \frac{1}{T-1} \sum_{i=2}^{T} Jaccard(q_{i-1}, q_T);$$

$$T3 = Lev(q_{T-1}, q_T) / \frac{1}{T-1} \sum_{i=2}^{T} Lev(q_{i-1}, q_i);$$

$$T4 = Lev(q_{T-1}, q_T) / \frac{1}{T-1} \sum_{i=2}^{T} Lev(q_{i-1}, q_T);$$

$$T5 = (t_T - t_{T-1}) / \frac{1}{T-1} \sum_{i=2}^{T} (t_i - t_{i-1})$$

5.2 Experimental Setups

Baseline methods we used include: 1) random prediction, 2) maximum category, 3) multilayer perceptron (MLP), 4) GBDT [24] and 5) XGBoost [4]. Among them, the maximum category is the most frequent class of each task in our dataset. Moreover, we implemented MLP with a two-layer neural network based on Pytorch [3]. As we mainly focus on the in-depth investigation of users' reformulating behaviors, how to design more sophisticated frameworks is beyond the scope of this paper.

We trained all supervised learning methods based on the extracted features and report their prediction performances in different tasks with 5-fold cross-validation. Note that in the "*Whether*" problem, we omitted the feature C2 as it could not be obtained before prediction. AUC and macro F1-score are used to evaluate

[3] https://pytorch.org

Table 5: Performances of all methods across different intent-level conditions in each task.

(a) The reason why users reformulate a query (Macro F1).

Model	IN	IT	TA	CL	BR	All
Random	0.115	0.143	0.163	0.130	0.093	0.192
Max Category	0.219	0.223	0.201	0.216	0.229	0.217
MLP	0.334	0.315	0.305	0.329	0.463	0.360
GBDT	0.387	0.343	0.319	0.375	0.589	0.384
XGBoost	**0.447**	**0.375**	**0.338**	**0.437**	**0.592**	**0.447**

(b) Whether users will access other reformulating entries (AUC).

Model	IN	IT	TA	CL	BR	All
Random	0.509	0.459	0.514	0.492	0.511	0.500
Max Category	0.5	0.5	0.5	0.5	0.5	0.5
MLP	0.786	**0.778**	0.716	0.746	0.714	0.759
GBDT	0.819	0.754	**0.735**	**0.769**	**0.688**	**0.776**
XGBoost	**0.819**	0.770	**0.735**	0.766	**0.688**	**0.776**

(c) Which reformulating entry users will access (Macro F1).

Model	IN	IT	TA	CL	BR	All
Random	0.147	0.139	0.139	0.157	0.238	0.152
Max Category	0.273	0.460	0.220	0.201	0.463	0.190
MLP	0.566	0.507	0.510	0.432	0.568	0.434
GBDT	**0.626**	0.622	**0.648**	**0.599**	**0.675**	**0.549**
XGBoost	0.536	**0.654**	0.568	0.538	0.657	0.538

* IN: interest-driven, IT: interest- and task-driven, TA: task-driven; CL: clear intent, BR: broad intent.

the predicting performance of the binary classification task. As for multi-class classification, we select macro F1-score and accuracy (ACC) as the evaluation metrics.

5.3 Results

In this section, we compare the performance of each classifier and each feature group. We further present the importance of each feature in prediction to serve as the guidance in these novel problems.

5.3.1 Comparison between classifiers. Table 5 reports the performance of each method in different tasks. We find that XGBoost performs the best in three tasks overall, followed by GBDT. All supervised learning methods achieve significantly better results than simple decisions, showing that it is possible to learn detailed user reformulating patterns. Most models predict why users reformulate a query more accurately in broad-intent sessions but have a poor performance on task-driven ones. Complex user behaviors in task-driven sessions may cause great difficulty for predicting user intents when they are leaving the current query (random decisions even achieves a higher F1 score value in the TA condition than the other two conditions). While in the BR condition, users tend to be satisfied as their search goals are easier to reach (i.e., Maximum Category outperforms Random), hence all methods can model user intents better. As revealed in Table 5(b), all learning methods have similar performances (decision trees are slightly better), indicating both linear and non-linear approaches are appropriate to predict whether users will access other reformulating entries. Different from Table 5(a), they can predict more accurately in the CL condition, where users prefer not to use the search box. As for the "*Which*" problem, decision trees are significantly superior to simple neural networks.

5.3.2 Comparison between feature groups. We further compare the performance of each single feature group using the best models in the corresponding task ("*Why*" & "*Whether*": XGBoost, "*Which*": GBDT), as shown in Table 6. In the "*Why*" and "*Which*" problems, query and mouse-related features perform the best among all groups, while in the "*Whether*" problem, the query group performs substantially better than others. Using all features yields the best performances in all metrics across various predicting tasks, suggesting the usefulness of each single feature group. The query group is overall the best. which indicates that exploiting more session contexts can improve the prediction of detailed aspects in reformulations.

5.3.3 Feature importance. To further verify the effectiveness of each feature, we inspect the importance (based on information gain) of them in these tasks [4], as shown in Figure 8, 9, and 10, respectively. Note that all importance scores have been normalized according to the maximum value. There are huge differences in the distribution of feature importance across three tasks. We find that in the "*Why*" problem, the top five most important features are the minimum/maximum clicked rank in the current query (C6/C7), the proportion of unique terms in previous queries (Q2), the number of previous queries (Q1), and the maximum browse depth in the current SERP (M7). All feature groups have a certain contribution to the overall system performance. While in the "*Whether*" problem, query related features contribute significantly more than other groups. Whether users will depend on themselves to reformulate queries may be highly related to their intents hence the query context will be effective. The number of clicks on other components in previous SERPs can also serve as a prospect, as users are more likely to use these interfaces if there are similar behaviors in former queries. This suggests the potential of introducing more personalized factors. Finally, mouse features are proved to be essential in predicting the interface that users adopt for reformulation, while other features manifest as trivial. Even so, predicting the entries that users access is meaningful because they reveal various directions of users' shifting intents.

To answer **RQ3**, we conclude that it is possible to predict finer-grained user intents behind their reformulations by leveraging session-level contextual information. Although our experiment is a premier step, it can serve as a reference for more intensive work in the future.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we have conducted a long-term field study to collect users' daily search activities as well as the fine-grained information of query reformulations. To our best knowledge, this is the first work to study multiple aspects of users' reformulating behaviors. Based on the collected data, we have thoroughly analyzed the distribution of reformulation types, reasons, inspiration source, and the entries that users access over time within sessions. To investigate the relationships between 1) user reformulations, 2) search effort & gain, 3) user actions and A) search intents as well as B) domain expertise, we have also made a detailed univariate analysis. Inspired by the findings in the field study, a supervised learning framework has been proposed to predict: 1) why users reformulate queries, and 2) how they reformulate these queries (i.e., from which entry).

[4] We conducted a feature analysis rather than the ablation study to report more detailed effectiveness of each feature.

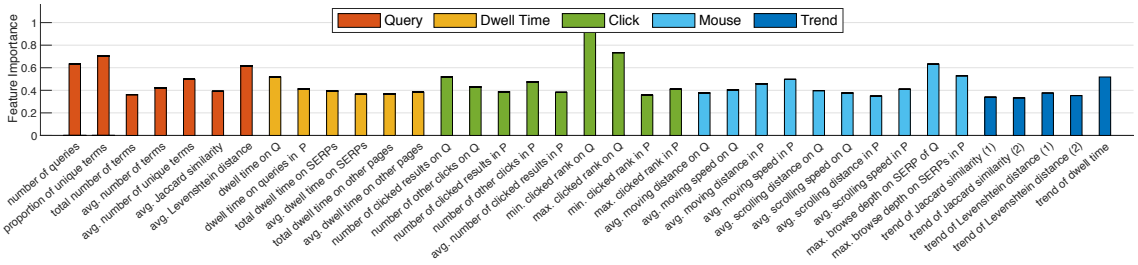


Figure 8: Normalized feature importance of predicting the reason behind each reformulation behavior using XGBoost.

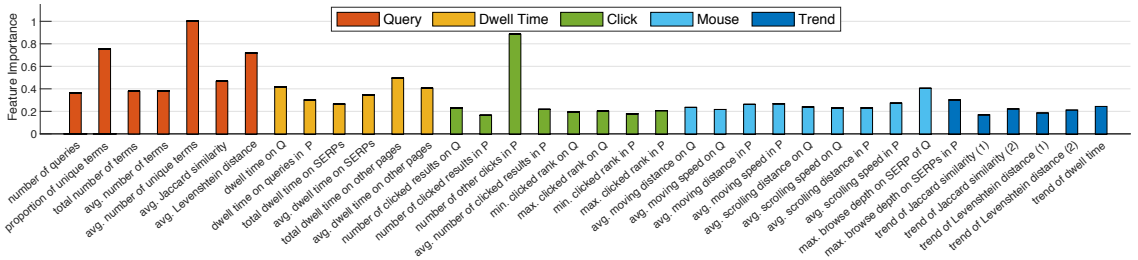


Figure 9: Normalized feature importance (w/o C2) of predicting whether users will access other entries using XGBoost.

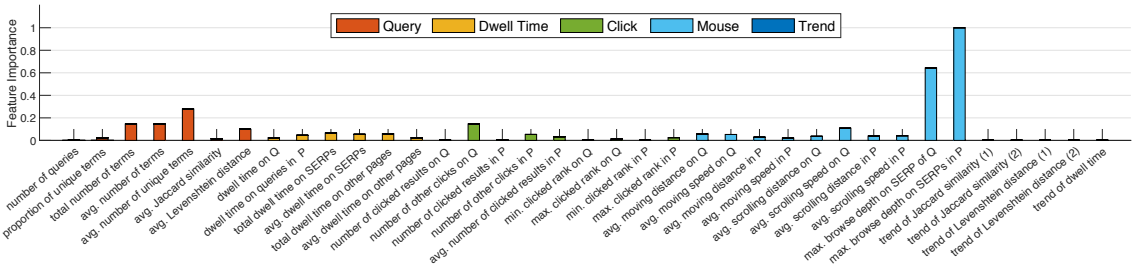


Figure 10: Normalized feature importance of predicting the entries that users access for query reformulation using GBDT.

Table 6: Comparison in predicting performance of single feature group in different tasks.

Feature Group	*Why*		*Whether*		*Which*	
	ACC	Macro F1	AUC	Macro F1	ACC	Macro F1
Query	0.755	0.257	0.682	0.718	0.736	0.465
Dwell	0.738	0.257	0.520	0.496	0.630	0.208
Click	0.750	0.285	0.635	0.669	0.745	0.418
Mouse	0.731	0.348	0.524	0.500	0.812	0.436
Trend	0.739	0.249	0.519	0.490	0.623	0.226
All	**0.774**	**0.444**	**0.776**	**0.814**	**0.839**	**0.549**

Our findings provide new insight into users' complex reformulating behaviors as well as the guidance for designing better query suggestion techniques in search engines. First, through the investigation of users' evolving query reformulations, we find that generally a search session can be summarized into a two-phase process: *Specialization→Intent Shift*. This suggests that the search scope of query recommendations can be broad at the beginning of sessions, but should be gradually narrowed down according to the follow-up queries issued by users. Second, existing search engines can better guide users in interest-driven tasks. However, their assistance in saving users' search efforts in complex or task-driven sessions is limited. As users pay more effort yet gain less satisfaction in complex tasks, search engines should be further improved to provide more help in these scenarios. One possible measure is to provide interactive summarizations (snippets) in the exploratory search process. Third, users behave differently in query reformulation with various search intent triggers and specificity. By understanding users' reformulating behaviors, we can better identify user intents and task property to further predict task difficulty as well as satisfaction. All of these are helpful in the optimization of search engines, e.g., balancing exploration and exploitation for recalling relevant documents or providing guidance for search users. Last but not least, according to the results in predicting why and how users reformulate a query, we find it possible to model delicate aspects of user reformulations by leveraging contextual information within sessions. Although this is the first attempt, our work will be beneficial for building more realistic user simulators in the future.

As with any research, there are limitations to our experiments: 1) Due to the limit of our browser extension, we only consider the reformulation entries in two commercial search engines. There may be more forms of reformulation interfaces in other search engines that we have not considered. 2) The reformulation inspirations were explicitly annotated by participants in the reviewing phase. However, sometimes they may not remember exactly which component has inspired them. More sophisticated techniques such as eye-tracking can be applied to collect more accurate information.

For future work, one possible direction is personalized query reformulation. In this paper, we mainly consider user reformulating behaviors within a search task. As different individuals may have different propensities or habits while using search engines, more information beyond sessions such as the search history can be considered to better understand user behaviors. We believe that such studies are forward-looking and can provide more insights into the design of SERPs.

REFERENCES

[1] Huanhuan Cao, Daxin Jiang, Jian Pei, Enhong Chen, and Hang Li. 2009. Towards context-aware search by learning a very large variable length hidden markov model from search logs. In *Proceedings of the 18th international conference on World wide web*. ACM, 191–200.

[2] Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating Query Reformulation Behavior of Search Users. In *China Conference on Information Retrieval*. Springer, 39–51.

[3] Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. TianGong-ST: A New Dataset with Large-scale Refined Real-world Web Search Sessions. In *Proceedings of the 28th ACM International on Conference on Information and Knowledge Management*. ACM.

[4] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[5] Wanyu Chen, Fei Cai, Honghui Chen, and Maarten de Rijke. 2017. Personalized query suggestion diversification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 817–820.

[6] Van Dang and Bruce W Croft. 2010. Query reformulation using anchor text. In *Proceedings of the third ACM international conference on Web search and data mining*. 41–50.

[7] Mostafa Dehghani, Sascha Rothe, Enrique Alfonseca, and Pascal Fleury. 2017. Learning to attend, copy, and generate for session-based query suggestion. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 1747–1756.

[8] Olive Jean Dunn. 1964. Multiple comparisons using rank sums. *Technometrics* 6, 3 (1964), 241–252.

[9] Carsten Eickhoff, Sebastian Dungs, and Vu Tran. 2015. An eye-tracking study of query reformulation. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 13–22.

[10] Bruno M Fonseca, Paulo Braz Golgher, Edleno Silva de Moura, and Nivio Ziviani. 2003. Using association rules to discover search engines related queries. In *Proceedings of the IEEE/LEOS 3rd International Conference on Numerical Simulation of Semiconductor Optoelectronic Devices (IEEE Cat. No. 03EX726)*. IEEE, 66–71.

[11] Wei Gao, Cheng Niu, Jian-Yun Nie, Ming Zhou, Kam-Fai Wong, and Hsiao-Wuen Hon. 2010. Exploiting query logs for cross-lingual query suggestions. *ACM Transactions on Information Systems (TOIS)* 28, 2 (2010), 1–33.

[12] Ahmed Hassan, Xiaolin Shi, Nick Craswell, and Bill Ramsey. 2013. Beyond clicks: query reformulation as a predictor of search satisfaction. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2019–2028.

[13] Jiyin He and Emine Yilmaz. 2017. 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*. 67–76.

[14] Sharon Hirsch, Ido Guy, Alexander Nus, Arnon Dagan, and Oren Kurland. 2020. Query reformulation in E-commerce search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1319–1328.

[15] Chien-Kang Huang, Lee-Feng Chien, and Yen-Jen Oyang. 2003. Relevant term suggestion in interactive web search based on contextual information in query session logs. *Journal of the American Society for Information Science and Technology* 54, 7 (2003), 638–649.

[16] Jeff Huang and Efthimis N Efthimiadis. 2009. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 77–86.

[17] Bernard J Jansen, Danielle L Booth, and Amanda Spink. 2009. Patterns of query reformulation during Web searching. *Journal of the american society for information science and technology* 60, 7 (2009), 1358–1371.

[18] Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W White. 2015. Understanding and predicting graded search satisfaction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. 57–66.

[19] Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, browsing, and clicking in a search session: changes in user behavior by task and over time. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 607–616.

[20] Jiepu Jiang and Chaoqun Ni. 2016. What affects word changes in query reformulation during a task-based search session?. In *Proceedings of the 2016 ACM on conference on human information interaction and retrieval*. 111–120.

[21] Jyun-Yu Jiang, Yen-Yu Ke, Pao-Yu Chien, and Pu-Jen Cheng. 2014. Learning user reformulation behavior for query auto-completion. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 445–454.

[22] Jyun-Yu Jiang and Wei Wang. 2018. RIN: Reformulation Inference Network for Context-Aware Query Suggestion. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 197–206.

[23] Lauri Kangassalo, Michiel Spapé, Giulio Jacucci, and Tuukka Ruotsalo. 2019. Why do users issue good queries? neural correlates of term specificity. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 375–384.

[24] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*. 3146–3154.

[25] Elad Kravi, Ido Guy, Avihai Mejer, David Carmel, Yoelle Maarek, Dan Pelleg, and Gilad Tsur. 2016. One query, many clicks: Analysis of queries with multiple clicks by the same user. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 1423–1432.

[26] William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* 47, 260 (1952), 583–621.

[27] Jiaxin Mao, Yiqun Liu, Noriko Kando, Min Zhang, and Shaoping Ma. 2018. How Does Domain Expertise Affect Users' Search Interaction and Outcome in Exploratory Search? *ACM Transactions on Information Systems (TOIS)* 36, 4 (2018), 1–30.

[28] Bhaskar Mitra. 2015. Exploring session context using distributed representations of queries and reformulations. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 3–12.

[29] Umut Ozertem, Olivier Chapelle, Pinar Donmez, and Emre Velipasaoglu. 2012. Learning to suggest: a machine learning framework for ranking query suggestions. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 25–34.

[30] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st international conference on Scalable information systems*. 1–es.

[31] Filip Radlinski, Martin Szummer, and Nick Craswell. 2010. Inferring query intent from reformulations and clicks. In *Proceedings of the 19th international conference on World wide web*. 1171–1172.

[32] Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. 2013. Learning to rank query suggestions for adhoc and diversity search. *Information Retrieval* 16, 4 (2013), 429–451.

[33] Philip Sedgwick. 2012. Multiple significance tests: the Bonferroni correction. *Bmj* 344 (2012), e509.

[34] Milad Shokouhi. 2013. Learning to personalize query auto-completion. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 103–112.

[35] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. 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*. ACM, 553–562.

[36] Xuanhui Wang and ChengXiang Zhai. 2008. Mining term association patterns from search logs for effective query reformulation. In *Proceedings of the 17th ACM conference on Information and knowledge management*. 479–488.

[37] Zhijing Wu, Yiqun Liu, Qianfan Zhang, Kailu Wu, Min Zhang, and Shaoping Ma. 2019. The influence of image search intents on user behavior and satisfaction. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 645–653.

[38] Fan Zhang, Jiaxin Mao, Yiqun Liu, Xiaohui Xie, Weizhi Ma, Min Zhang, and Shaoping Ma. 2020. Models Versus Satisfaction: Towards a Better Understanding of Evaluation Metrics. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 379–388.

[39] Xiangmin Zhang, Hermina GB Anghelescu, and Xiaojun Yuan. 2005. Domain Knowledge, Search Behaviour, and Search Effectiveness of Engineering and Science Students: An Exploratory Study. *Information Research: An International Electronic Journal* 10, 2 (2005), n2.