Searching, Browsing, and Clicking in a Search Session: Changes in User Behavior by Task and Over Time

Jiepu Jiang[1], Daqing He[2], James Allan[1]

[1] Center for Intelligent Information Retrieval,
School of Computer Science, University of Massachusetts Amherst

[2] School of Information Sciences, University of Pittsburgh

jpjiang@cs.umass.edu, dah44@pitt.edu, allan@cs.umass.edu

ABSTRACT

There are many existing studies of user behavior in simple tasks (e.g., navigational and informational search) within a short duration of 1–2 queries. However, we know relatively little about user behavior, especially browsing and clicking behavior, for longer search session solving complex search tasks. In this paper, we characterize and compare user behavior in relatively long search sessions (10 minutes; about 5 queries) for search tasks of four different types. The tasks differ in two dimensions: (1) the user is locating facts or is pursuing intellectual understanding of a topic; (2) the user has a specific task goal or has an ill-defined and undeveloped goal. We analyze how search behavior as well as browsing and clicking patterns change during a search session in these different tasks. Our results indicate that user behavior in the four types of tasks differ in various aspects, including search activeness, browsing style, clicking strategy, and query reformulation. As a search session progresses, we note that users shift their interests to focus less on the top results but more on results ranked at lower positions in browsing. We also found that results eventually become less and less attractive for the users. The reasons vary and include downgraded search performance of query, decreased novelty of search results, and decaying persistence of users in browsing. Our study highlights the lack of long session support in existing search engines and suggests different strategies of supporting longer sessions according to different task types.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process.* H.1.2 [**Models and Principles**]: User/Machine Systems – *human factors.*

General Terms

Experimentation, Human Factors.

Keywords

Session; task; search behavior; browsing; clicking; eye-tracking.

1. INTRODUCTION

Although some simple search problems (e.g., finding a specific homepage and locating specific facts with known keywords) can be satisfied through a single query and one click, it usually takes multiple searches to solve more complex tasks. The reasons vary. Sometimes it is the user who adopts a divide and conquer strategy, using each query to deal with a part of the task's goal [1]. Also sometimes it may be the complexity of the solution that makes it difficult to find all the answers with one query. Moreover, the user usually does not start with a clear goal and needs to figure out a specific information need after many searches [30]. For whichever reason, a search process that solves a complex problem usually spans more than one query and includes rich user interaction.

Studies of users' search behavior provide guidance to system design and evaluation. With many studies of user behavior in simple search tasks (1–2 queries), we know relatively well how to tailor a system for these tasks. For example, after Joachims et al. [15] showed that users' visual attention and clicks are biased to the top ranked results, we know systems achieving high precision are more preferable in web search than systems with high recall.

In comparison, we know relatively little about user behavior in long sessions of complex task types, especially those that can provide guidance to the design and evaluation of systems supporting long session and complex tasks. For example, do users examine more result snippets and go to deeper ranks in complex tasks and long sessions? Are users looking for factual information more accurate in clicking given short result snippets? Do users become less persistent in viewing the search engine result page (SERP) after long durations of search? Without knowing answers to these questions, we do not know how to design and evaluate systems to better support complex tasks and long search sessions.

To address that gap, we conducted an experiment with users working on complex tasks for relatively long search session (10 minutes; about 5 queries) and recorded search behaviors including eye movement data. We study the following research questions:

RQ1: How do users' search behaviors, especially browsing and clicking behaviors, vary in complex tasks of different types?

Studying this question helps us understand the effects of tasks on personalization and suggests how to design systems supporting complex tasks. To the best of our knowledge, among web search user studies focusing on SERP browsing patterns, our experiments involve the most complex tasks and the longest sessions. Joachims et al.'s experiments [15, 23] dealt with only navigational and informational tasks, with on average 1.6 queries issued per session. Moffat et al. [24, 29] did not report session length, but according to Wu et al. [31] the most complex tasks adopted by Moffat et al. involve 2.42 queries and 3.46 clicks. Cole et al. [6, 7, 21] included tasks comparably complex to our study, but they focused on how users shift between scanning and reading.

We will show that there are very noticeable differences of behavior depending on the type of task driving the user's search. Differences are present in how active users are, how they browse and click result abstracts in a SERP, and how they issue queries.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609633

RQ2: How do users' search behaviors change over time in a search session?

Our study is also the first study with analysis of changes in SERP browsing and clicking patterns over time in relatively long search sessions (10 minutes). Answers to RQ2 may provide insights on how to support users in long sessions. We will show that user engagement with a search system changes substantially between the start and the end of these search sessions. The changes appear to largely reflect a loss of confidence in the results, along with shifted patterns in browsing SERP results.

The rest of this paper introduces our experiments and findings.

2. RELATED WORK

Our study is related to three areas of existing work: web search user behavior with eye-tracking data; search task and its effects to user behavior; and search session. We review each area below.

Early studies (before 2004) of web search user behavior are mostly based on the analysis of large-scale query logs, such as [13, 28]. These studies described what real life web searches are like and how users interact with search engines at the query level, but they do not provide details of user behavior on a SERP, such as how users examine result abstracts. The first work using eye-tracking for web search user behavior [11, 15, 23] discovered how users browse a SERP, examine abstracts, and click results. They found decayed visual attention on results as the rank of the result increases and biased clicks on the top ranked results. Lorigo et al. showed that task type and gender may result in differences in search behavior and browsing style [23]. Later studies with eye-tracking [3, 8, 10, 29] further confirmed that users behave diversely in different tasks. They also showed that users may react distinctly to different outlooks of search result abstracts [5, 8] and SERP elements other than result abstracts, such as ads and related searches [3, 10]. More recently [24] used eye tracking studies to verify models and hypotheses in IR evaluation metrics. However, due to the limited accessibility of devices, user behavior studies with eye-tracking data are limited.

Although currently lots of work using eye movement data for user behavior studies exist, the search tasks being studied in [3, 8, 10, 11, 15, 23] are simple, e.g., the "navigational" and "informational" tasks defined in [2]. As reported in [23], on average 1.6 queries were issued in that work. Recently Moffat et al. [24, 29] used more complex tasks of different cognitive complexity, i.e., "remember", "understand", and "analyze" defined in [31]. However, even the most complex task type ("analyze") only involves search sessions of 2.42 queries and 3.46 clicks on average [31]. To the best of our knowledge, among existing search behavior user studies with eye-tracking data, only [6, 7, 21] conducted experiments based on tasks of comparable complexity to the tasks adopted in our paper. However, they did not study how users read result abstracts in a SERP, but focused on how they shift between scanning and reading [6, 7, 21]. Therefore, it is unclear how users behave—especially how they browse the result abstracts in a SERP and click results—in long sessions of complex tasks.

When tasks are complex, it usually requires relatively longer search sessions to finish. Previous studies using web search logs [13, 28] discussed search sessions as multiple searches across certain duration of time in search logs. However, from this aspect, the multiple searches within a session are not necessarily related to a consistent topic or search task. Spink et al. [27] found that multi-tasking is very common in search sessions derived using this definition. In our study, a search session refers to consecutive searches that aim at solving a consistent task, which is similar to the search sessions studied in [6, 7, 20–22].

3. EXPERIMENTS

We conducted an experiment to collect user behavior data in search sessions for completing complex tasks. We collected users' queries, their browsing of Search Engine Result Pages (SERPs) and their clicks of search results. In addition, we deliberately asked users to perform different types of search tasks.

3.1 Search Tasks

The search tasks involved in previous related studies mainly focused on short search sessions and mostly dealt with navigational and simple informational needs. For example, in Joachims et al. [15] and Lorigo et al.'s studies [23], users on average only issued 1.6 queries in each task, and the tasks in Cutrell and Guan's studies [8, 12] were simplified so that both navigational and informational search tasks were considered to be successful once a best result page was found. Our study of search behaviors, particularly the examination of changes over time, needs to work on relatively longer search sessions, which allow a long exploration process and complex user interactions. We therefore adopted search tasks from the TREC 2012 session track [17], which were categorized into four types using Li and Belkin's faceted classification approach [19].

We considered two facets of search tasks identified by Li and Belkin [19]: product and goal. The product of a search task can be either *factual* (to locate facts) or *intellectual* (to enhance the user's understanding of a topic). The goal of a search task can be either *specific* (well-defined and fully developed) or *amorphous* (ill-defined or unclear goals that may evolve along with the user's exploration). This yields four types of tasks: known item search (KI), known subject search (KS), interpretive search (IN), and exploratory search (EX). Some examples of tasks are:

Known Item (factual + specific): *Where is Bollywood located? From what foreign city did Bollywood derive its name? What is the Bollywood equivalent of Beverly Hills? What is Bollywood's equivalent of the Oscars? Where does Bollywood rank in the world's film industries? Who are some of the Bollywood stars?*

Known Subject (factual + amorphous): *You think that one of your friends may have depression, and you want to search information about the depression symptoms and possible treatments.*

Interpretive (intellectual + specific): *You would like to buy a dehumidifier. You want to know what makes a dehumidifier good value for money.*

Exploratory (intellectual + amorphous): *You would like to buy a dehumidifier. On what basis should you compare different dehumidifiers?*

Although these four types of tasks appear to be different from those tasks presented in previous works [8, 12, 15, 23] (navigational and information search tasks), we believe that the two classification schemes do not conflict with each other, but are defined at different levels. Broder defined navigational and informational search tasks [2] based on a classification of individual web search queries. Therefore, each task in this case is intrinsically only indicative of what people can finish within a single query. In comparison, Li and Belkin's classification scheme [19] is defined regarding the nature of people's information needs and problems, and allows multiple queries in a search session. Of course, each query in the session may still fit into Broder's scheme [2]. For example, in IN and EX tasks, users may issue a navigational query "amazon" to know about the different types of dehumidifiers sold on amazon.com.

3.2 System

We built an experimental search system providing modified Google search results. First, all ads and sponsors' links were removed. Second, we showed 9 results each page (rather than the usual 10) to make sure that users do not need to scroll down to see

all of the result items. This change made it much simpler to analyze eye-tracking data. However, previous studies [15] also reported that scrolling down affected browsing patterns on results shown below the screen cutoff of a search result page. This change will miss such effects. We adopted this change because Joachims et al. [15] showed that most of the users' attention is still focused on the top ranked results which are visible before scrolling down. Third, if Google provided query suggestions (i.e., "related searches") for a query (usually shown below the search results), we moved them to the right side of the search results, again, to eliminate scrolling pages.

The system looks very similar to existing search engines except a few places specifically designed for our search tasks. It shows the task descriptions at the top of the search result page. This is because we found in our pilot study that, without showing the task description, users might constantly switch between search result pages and another page showing the task description, because they forgot details of the task. We believe this would cause greater issues to the collected data (e.g., more constantly switching of pages) than showing task description on the search result page. In addition, the system has a highlighted "finish task" link if the session exceeds the time limit (but not before the limit is reached). Although many systems for user studies (e.g. Liu et al.'s systems [21]) allow users to bookmark relevant results at search time, we did not adopt such settings because it may affect users' browsing behaviors. Instead, relevance judgments were completed after search.

3.3 Eye-Tracking

A Tobbi 1750 eye-tracker was used to collect eye movement data. Among the various types of eye movement data, we only focus on analyzing fixation: stably gazing at an area of the screen. Studies have shown that fixation on an area of the screen usually indicates that users are reading information displayed on the area of interest (AOI) [26]. The AOIs in our study include each search result abstract (snippet), query suggestion, and task description. We assume that fixation on these AOIs indicate that the participant has examined the corresponding result abstract, query suggestion, and task description. ClearView, a software accompanying the eye-tracker, was used to analyze fixations on the defined AOIs. We set the minimum duration of fixation to 100ms, a common value adopted in many previous studies of web search behaviors using the same series of eye-tracker [8, 12].

In the following discussion, we say that the participant examined a result abstract if we observed fixations on its corresponding AOI when the participant was browsing search result pages. Similarly, we say that the participant examined the query suggestions or topic description if we observed fixations on the corresponding AOIs.

3.4 Participants

We recruited 20 English native speakers (13 female and 7 male) through flyers in the campus of University of Pittsburgh. We required the participants to be English native speakers and current students in a college or university. Considering previous studies [9] reported increased error rates of eye-tracking for participants wearing glasses or lens, we also specified in our ads that the participants should have perfect eyesight (20/25). 13 participants were aged between 25–30, 6 between 18–24, and one over 30. For the highest degree earned or expected, 9 reported bachelor degree, 9 master, and 2 doctoral. Eight participants were studying information related majors, while others' majors ranged from anthropology to microbiology. The participants rated their expertise of using web search engines by a 5 point Likert scale and the mean score is 3.75 (5 means the highest proficiency).

3.5 Experiment Procedure

We randomly sampled five groups of search tasks developed by the TREC 2012 session track [17]. Each group has four unique tasks of different types. For each task group, four participants worked on the tasks. We rotated the sequence of tasks in each group for different participants. Table 1 shows the topics.

Table 1. Search task assignments.

Group	TREC Topic No. & Task Type	Participants
1	32 (KI), 40 (KS), 07 (IN), 05 (EX)	S01 – S04
2	11 (KI), 22 (KS), 02 (IN), 29 (EX)	S05 – S08
3	15 (KI), 03 (KS), 39 (IN), 10 (EX)	S09 – S12
4	30 (KI), 33 (KS), 41 (IN), 37 (EX)	S13 – S16
5	23 (KI), 04 (KS), 48 (IN), 46 (EX)	S17 – S20

The total experiment duration for a participant is about 2 hours. The participants were reimbursed by the rate of $15 per hour. At the beginning of the experiment, participants were introduced to the system and a training task (with all the three stages but shorter time limits). Then the participants worked on four formal tasks. After two formal tasks, they took a 10-minute break. We interviewed the participants for their search behaviors at the end of the experiment. For each task, the participants finished the following stages:

1. Search (10 minutes). In the search stage, the participants were introduced to the search task and were asked to use the experimental system to find information in order to solve the task. They were instructed to use the experiment system as if they were using public search engines such as Google and Bing—e.g. they could search using textual queries, browse search result pages, click and view results. However, they were specifically instructed not to use other search engines. We set a limit of 10 minutes for each task. After 10 minutes, the system showed a highlighted link notifying them to terminate the search stage. However, we also allowed participants to finish the task before 10 minutes if they reported they had already learned enough to solve the task.

2. Report (5 minutes). In the report stage, the participants were asked to rate the difficulty of the task, their familiarity with the topic of the task prior to search, and their search performance using a 5 point Likert scale. Then they were asked to write a paragraph reporting their outcomes of the search task. During this stage, the system showed a countdown of 5 minutes to help the participants to finish in about that time. The system did not freeze after 5 minutes. We instructed the participants to make full use of the time instead of finishing as soon as possible.

3. Relevance Judgments (5 minutes). In this stage, the participants were asked to judge and rate results regarding their relevance to the search task. Due to the time limit of the experiment, it was usually impossible to judge all returned results of all queries in a session. Therefore we generated a pool of results for relevance judgments. The priority of selecting results (from high to low) is: clicked results, probably examined results, other results.

The pool size was about 25 results. First, we included all the results that the user clicked on. If less than 25 results (say, N_c results) were clicked, we continued to consider some "probably examined" results. We assume the participants looked through each search result page from top to bottom. Therefore, we located the deepest position of the clicked results in each search result page and considered unclicked results ranked higher than the deepest position as "probably examined results". If there were no more than 25-N_c "probably examined results", we included all into the pool; or, we randomly sampled 25-N_c. If summing up all clicked results and probably examined results did not total 25, we further included a random sample of other results into the pool.

This pooling procedure is to maximize the number of judged results among those were clicked and examined throughout the session. The participants rated each result as "highly relevant",

"somewhat relevant", or "non-relevant". The system forced the participants to click each result at least once before submitting judgments. Again, the system showed a countdown of 5 minutes and the participants were instructed to make full use of the time. Later, an external annotator judged the rest of the results.

4. DATA

We collected user behaviors from 80 search sessions on 20 unique tasks. During a search session, the participants on average issued 4.9 queries, examined 16.1 unique result abstracts, and clicked 9.3 unique results. The average length of a query was 3.96 words (without removing stopwords). As with most search engines, if the participant clicks a result, the experiment system left the current search engine result page (SERP) and switched to a new tab of the browser showing the result webpage. The participants needed to switch between the SERP and result webpages. This resulted in multiple views for a SERP. We refer to the duration from showing a SERP to switching to other webpages as a "SERP view". In our experiment, participants had 3.6 views for a SERP on average.

For each session, the participant on average judged 20.1 results and left 13.3 unjudged. In total this resulted in 992 unique unjudged task-URL pairs. In order to evaluate search performance of sessions, we asked an annotator (not an author of this paper) to assess the relevance of the unjudged results. To evaluate the agreement between the annotator and the participants on relevance judgments, we also sampled 100 unique judged results for the annotator to assess. The annotator was not aware of which result has been judged by the participants. If we merge "highly relevant" and "somewhat relevant" into one class, the annotator agreed with the participants on 77% of the cases.

To evaluate the correctness of the fixation data, we calculated the percentage of clicked results with observed fixations. Intuitively, the user should have examined a result abstract before clicking it. Therefore, we should observe fixations on the clicked results if the data is accurate. In our experiment, the percentage is 87%, comparable to those reported by Joachims et al. [15].

5. SEARCH BEHAVIORS

Users interact with a search engine mainly in two ways. First, they proceed through the search process by issuing queries. Second, for each query, they examine result abstracts on the SERP and may click on results. Therefore, we first compare the "search activeness" of users in terms of how frequently they search and how often they examine and click results in Section 5.1. This helps us understand the diverse weight of the two types of interactions in different tasks. Then we look into details of SERP browsing in Section 5.2 and results clicking in Section 5.3. Finally, we compare users' querying reformulation behaviors in Section 5.4.

5.1 Search Activeness

Search activeness examines how active users are in terms of the frequency of query and result level interactions. Specifically, we compare: search frequency (# queries); the frequency of viewing SERPs (# SERP views); the number of examined result abstracts (# unique fixations) and clicked results (# unique clicks); total time of viewing a SERP or SERPs (% or # time view SERP). Table 2 shows results for a session and for an individual query (labeled with "/ q").

It should be noted that the length of a task session is not strictly 10 minutes. A session can be shorter if the user chooses to finish before the time limit is reached. It may also be longer than 10 minutes it the user does not realize the time is up. This is because the system only shows the notification on the search page, but the users may be reading result webpages when the time limit expires. As shown in Table 2 ("Total task time"), users spent about 10% longer in KS tasks, while the time of other tasks does not differ greatly. This is probably due to the fact that fewer users chose to finish a KS task before 10 minutes ("# sessions end by user").

As shown in the table, users in different tasks can be active in diverse ways. For example, users in KI and EX sessions tend to search more frequently but interact less actively in each search, while KS and IN sessions involve fewer searches in total but more activities during each search. Data in Table 2 shows that users issued 5.5 and 6.2 queries in KI and EX sessions, which is significantly more often than those in KS (4.2 queries) and IN tasks (3.6 queries) within roughly the same amount of time. However, during each individual search, users in KS and IN sessions viewed the SERP more frequently (2.96 and 3.41 SERP views) and clicked more search results (2.32 and 2.58 unique clicked results), which are significantly more active than they did in KI and EX sessions (2.17 and 2.38 SERP views; 1.58 and 1.96 unique clicks). In addition, users in KI, KS and IN sessions spent longer total duration (12.4s–13.9s) viewing a SERP than they did in EX sessions (10.7s).

Table 2. Users search activeness in different types of tasks.

Statistics	KI	KS	IN	EX	
Total task time (s)	599	**651**	600	581	KI<KS*, KS>IN*, KS > EX**
# sessions end by user	**4/20**	1/20	2/20	3/20	
# queries	5.5	4.2	3.6	**6.2**	KI>IN *, KS<EX +, IN<EX **
# SERP views	12.0	12.5	12.1	**14.7**	
# unique fixations	16.6	13.6	10.7	**18.6**	KI>IN*, KS<EX^{+}, IN<EX**
# unique clicks	8.3	9.5	8.2	**10.7**	KI<EX^{+}, IN<EX^{+}
% time view SERP	**22.0**	13.2	13.8	21.2	KI>KS**, KI>IN**, KS<EX**, IN<EX**
# SERP views / q	2.17	2.96	**3.41**	2.38	KI<KS*, KI<IN**, IN>EX**
# unique fixations / q	3.37	3.48	**3.75**	3.61	
# unique clicks / q	1.58	2.32	**2.58**	1.96	KI<KS**, KI<IN**, IN>EX*
Time view SERP / q	13.1	12.4	**13.9**	10.7	

$^{+}$, *, **: difference is significant at 0.1, 0.05, and 0.01 level.

According to Table 2, EX sessions are the most active among the four tasks in terms of the frequency of search and SERP views and the number of examined/clicked results in a search session. In comparison, users in KS and IN sessions are less active. They spent significantly shorter time on SERP views (13.2% and 13.8% of the session) than they did in KI and EX sessions (22.0% and 21.2%). They also examined fewer abstracts during the session (13.6 and 10.7 unique fixations. KI sessions are less active than EX in that users clicked fewer results, but KI sessions are more active than KS and IN because of more examined result abstracts and longer durations of viewing SERPs.

The diverse styles of search activeness suggest that we can support a task according to the popularity of query and result level interactions in the task (once we know what types of tasks users are dealing with). For example, in KI and EX tasks, we may support search sessions by assisting with query reformulation (because they search more often in a session). In comparison, with less query level interaction and more result level actions, KS and IN tasks should be supported by focusing on enhancing result level interaction. For example, search results for KS and IN tasks can be optimized for precision at lower ranks or of a whole page.

5.2 SERP Browsing

This section compares different tasks based on users' browsing styles. Results are compared from four aspects: the effort of a SERP view, the chances of examining results at different ranks, the sequence of examining result abstracts on a SERP (scan path), and the users' attentions on visited results.

5.2.1 Effort of SERP Browsing

We found that users in KI and EX sessions spent greater effort on examining result abstracts in a SERP view. As shown in Table 3, users examined significantly more unique abstracts in KI and EX tasks (2.48 and 2.59 unique fixations) than they did in KS and IN tasks (2.16 and 1.93). We further aggregated the durations of all the fixations on result abstracts in a SERP view ("Fix time on results").

Table 4. Statistics for users' scan path in a SERP view.

Statistics	KI	KS	IN	EX	
P(moving up)	0.31	0.29	**0.45**	0.37	KI<IN**, KS<IN**
P(sequential)	0.28	**0.41**	0.07	0.21	KI>IN**, KS>IN**, KS>EX*, IN<EX*
Breadth	3.09	3.44	3.17	**4.02**	KI<EX*
Gap	1.38	1.47	1.54	**1.63**	

*, **: difference is significant at 0.05 and 0.01 level.

It shows that in a SERP view, users in KI and EX tasks also used longer durations in total on examining result abstracts (1.94s and 1.91s) comparing to those in KS and IN tasks (1.35s and 1.46s). In addition, users in KI and EX tasks examined each single result abstract for longer periods than they did in KS tasks ("Fix time on a result").

In comparison, users in KS and IN sessions spent more time on reading result webpages. Between each SERP view, users can read and explore result webpages. Although users may follow links on the result webpage and visit new webpages, the time interval of two SERP views to some degree tells the amount of time the user spent on each result webpage. As shown in Table 3 ("SERP view interval"), it took a significantly longer time for users in KS and IN tasks to switch back from result webpages to SERP, probably indicating that they spent more time reading each result webpage. Table 2 also supports our conjecture. The percent of time users spent on viewing SERPs ("% time view SERP") is significantly less in KS and IN tasks.

Table 3. Browsing behavior statistics for a SERP view.

Statistics	KI	KS	IN	EX	
# unique fixations	2.48	2.16	1.93	**2.59**	KI>KS*, KI>IN**, KS<EX**, IN<EX**
Fix time on results	**1.94**	1.35	1.46	1.91	KI>KS**, KI>IN**, KS<EX**, IN<EX**
Fix time on a result	**0.75**	0.60	0.70	0.72	KI>KS**, KS<IN*, KS<EX**
Length a SERP view	**6.04**	4.17	4.07	4.50	KI>KS**, KI>IN**, KI>EX**
SERP view interval	28.6	**38.5**	36.1	21.9	KI<KS**, KI<IN*, KI>EX*, KS>EX**, IN>EX**
Avg examined rank	3.18	3.78	3.95	**3.96**	KI<KS**, KI<IN**, KI<EX**
Max examined rank	4.35	5.06	5.23	**5.36**	KI<KS**, KI<IN**, KI<EX**
% fixations on visited	21.6	22.4	**22.7**	21.7	

*, **: difference is significant at 0.05 and 0.01 level.

As users spent more effort examining result abstracts in KI and EX tasks, systems supporting these tasks should consider how to generate more informative result abstracts. In the same way that Cutrell et al. found that different lengths of result abstracts can affect performance of navigational and informational search differently [8], KI and EX tasks may also benefit from customized styles of result abstracts specifically trimmed for the tasks. In comparison, KS and IN tasks may benefit from various reading supports for the result webpages (e.g., highlighting query terms in a result webpage when users are redirected from a SERP).

5.2.2 Attention on Results at Different Ranks

Previous studies [15] showed that users focus more on top ranked results when examining a SERP. Does such tendency exist in a search session and is it different in the four types of tasks? Figure 1 shows the chances of examining results (fixation rates) at different ranks (we refer to the result at rank *n* as Rn). The left figure counts the fixation rates only for the first view of a SERP, while the right one aggregates all SERP views in a session. Both figures show that users still examined more on results at higher ranks in a search session, but vary slightly in patterns.

The results show that during the first view of a SERP, users in EX tasks were more willing to examine results at the bottom of a page comparing to other tasks. As shown in Figure 1 (left), users in EX sessions had 10%–20% chance of examining R7–R9, while this happened in less than 10% of the cases in other three tasks. However, when considering all SERP views of a query, KS, IN, and EX sessions have comparable fixation rates on lower ranked results (about 40%–50%), but KI sessions still showed observably lower tendency to examine results at the bottom (about 35%). This may be caused by the fact that users viewed a SERP more times in

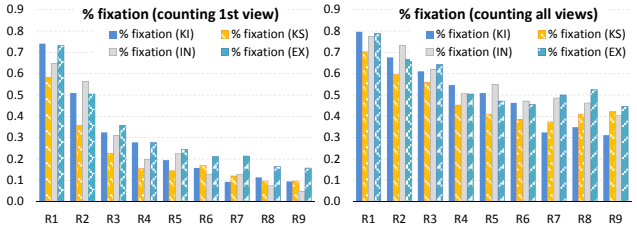

Figure 1. Fixation rates on results at each rank (R1–R9), counting the first view (left) or all views (right) of each SERP.

KS and IN sessions than they did in KI sessions (see Table 2). As found by Lorigo et al. [23], when viewing a SERP multiple times, users shift attention to focus more on results at the bottom. Therefore, if counting all SERP views, it is not surprising that users in KS and IN sessions may increase their fixation rates on R7–R9 after viewing a SERP multiple times.

Results in Figure 1 shows that, unlike simple search that may be satisfied with one click (and therefore one SERP view if the click is accurate), in complex search tasks, results at lower rank positions of a SERP can still get substantial visual attention (about 30%–50% fixation rate) after multiple SERP views. This suggests that it is unnecessary to rigorously optimize results for precision at the very top positions in long sessions of complex tasks. In addition, results suggest that tasks do affect fixation rates, and therefore systems can be tailored for different browsing styles.

5.2.3 Scan Path

Solely looking into fixation rates is often not indicative of how users consecutively examine result abstracts in a SERP view. Therefore, we study the users' "scan paths" in a SERP view. As Lorigo et al. did in their studies [23], we aggregate the examined results in a SERP view as a "scan path". If the users examined the same result abstract repeatedly, we count the result only once in the scan path. For example, for an observed sequence "R1 R3 R3 R1 R4 R4", its corresponding scan path is "R1 R3 R1 R4".

We refer to two adjacent examined abstracts in a scan path as a move. For example, R1–R3, R3–R1, and R1–R4 are three moves in "R1 R3 R1 R4". The distance of the two results in a move is referred to as a "gap", and we define the gap of a scan path as the average gap of all its moves (e.g., "R1 R3 R1 R4" has gap 2, 2, 3 and its average gap is 2.33). The gap of a scan path can indicate how many results are skipped in browsing. A scan path with gap 1 means that each move is going to an adjacent result. We define the "breadth" of a scan path as the maximum gap of two examined result abstracts in the scan path (e.g., the breadth of "R1 R3 R1 R4" is 3, the distance of R1 and R4). The breadth of a scan path can indicate the magnitude of the area being browsed in a SERP. If the users examine results from top to bottom sequentially, each move in the scan path would be "moving down" (to the results at lower ranks). If all the moves in a scan path are "moving down", we say that the scan path is "sequential" (from top to bottom). We estimate the chances of "moving up" and the chances of a scan path being sequential. Table 4 shows the results.

Table 4 suggests that instead of scanning the whole page, users focus on an area of 3–4 results in a SERP view ("breadth") and usually skip results in browsing ("gap" ranges from 1.38 to 1.63). Although Table 3 shows a comparable amount of fixations in KI and EX tasks, Table 4 explains the difference between KI and EX. In EX sessions, users' scan paths have wider breadths (4.02) than those in KI sessions (3.09). This indicates that users in EX tasks browsed larger areas and skipped more results in a SERP view, while users focused on smaller areas in KI sessions.

As shown in Table 4, in all tasks the chance of moving up is lower than 50%, showing that users tend to scan results in a SERP

from top to bottom in general. However, the chances of going up is significantly lower in KI and KS tasks (about 30%) compared to those in IN tasks (45%). Note that 45% chance of moving up means that users in IN sessions are almost randomly moving toward either the top or the bottom. Thus it is not surprising that only 7% of the scan paths in IN tasks are sequential.

Table 4 also show strong dimensional characteristics. Tasks looking for factual products (KI and KS) show significantly stronger tendencies of sequential browsing ($p<0.01$). Tasks with amorphous goals (KS and EX) have significantly greater gap and breadth in a SERP view ($p<0.01$). This indicates that in both dimensions, more complex tasks (e.g. informational product and amorphous goal) lead to more complex browsing behaviors – e.g. non-linear browsing and scanning larger areas.

5.3 Results Clicking

We further compare the four types of tasks based on the users' clicking behaviors. Whether or not a result is clicked depends on two factors. First, whether the user examined the result abstract or not (though possible, it is very unlikely that users blindly open a result without examining it). The previous section examined that factor. This section focuses on the second one: after examining a result, what is the chance a user clicks on it? The results show that users do not click every result abstract they examined. The chance of clicking varies by task, by relevance of results, and by whether the result has been visited previously.

5.3.1 Chances of Clicking Examined Results

We calculate the probability of clicking a result provided that we observe a fixation on the result abstract during a SERP view. Table 5 shows the results – "P(click | examine)". It shows that users are significantly more likely to click a result after examining it in KS and IN sessions (61% and 59%), whereas the chances are lower in KI and EX sessions (45% and 52%). This is not surprising considering the fact that users in KI and EX tasks also retrieved fewer relevant results. As shown in Table 6 (analyzed in greater detail in Section 5.4), P@10 and nDCG@10 in KI and EX sessions are significantly lower than those in KS and IN sessions. With fewer relevant results retrieved, the examined results are less likely to be relevant and therefore less likely being viewed as promising and so worth clicking by the users.

We also noticed that users do not always click results during a SERP view. Sometimes they switch from a result webpage to the SERP and then switch back to the result webpage again, probably because they did not find any interesting results after examining the SERP. The chance of viewing a SERP without clicking result ("% SERP views w/o clicks") is lower in EX sessions. Users also clicked significantly more results in EX tasks during a SERP view (0.87 unique clicks) comparing to other tasks (0.77–0.80).

5.3.2 Relevance of Results and Clicking

To evaluate how relevance of results affects a user's decision to click in the four types of tasks, we further calculate the chance of clicking an examined result abstract when the result is judged as relevant (either "highly relevant" or "somewhat relevant"). As shown in Table 5, P(click | examine, relevant), the chance of clicking increases by 6%–16% if the examined result abstract is relevant. When a relevant result abstract has been examined, users in KS, IN, and EX sessions have comparable chances of clicking the result (65%–70%). However, users in KI sessions still have significantly lower chances of clicking (56%). This may indicate that users intrinsically click more selectively in KI tasks.

Table 5. Clicking behavior statistics in a search session.

Statistics	KI	KS	IN	EX	
# unique clicks / SERP view	0.77	0.80	0.79	**0.87**	KI<EX**, KS<EX*, IN<EX**
% SERP views w/o clicks	17.0	**21.0**	19.0	13.5	KS>EX*
P(click | examine)	0.45	**0.61**	0.59	0.52	KI<KS**, KI<IN**
P(click | examine, relevant)	0.56	**0.70**	0.65	0.68	KI<KS**, KI<IN**, KI<EX**
P(click | examine, visited)	0.36	0.33	0.39	**0.44**	
% clicks relevant	**87.7**	87.2	79.6	73.2	KI>IN*, KI>EX**, KS>IN*, KS>EX**
% clicks visited	9.2	2.7	11.1	**14.6**	KI>KS**, KI<EX*, KS<IN**, KS<EX**
Avg clicked rank	2.94	3.51	**3.72**	3.46	KI<KS*, KI<IN**, KI<EX*
Deepest clicked rank	4.15	5.41	**5.58**	4.78	KI<KS**, KI<IN**, IN>EX*

*, **: difference is significant at 0.05 and 0.01 level.

Unsurprisingly, users cannot perfectly predict whether a result is useful or not purely based on the abstract returned by a search engine. As shown in Table 5 ("% click relevant"), the percentage of relevant results among all clicked results varies from task to task: over 87% clicked results in KI and KS tasks are relevant, which is a significantly higher percentage than those in IN and EX tasks. This may also indicate that it is easier for users to judge the usefulness of documents if they are searching with a factual goal.

The lower click accuracy in tasks looking for intellectual products (IN and EX tasks) indicates that the result abstracts provided in current search engines are probably optimized for factual search only, which is difficult to satisfy users searching for other types of information. Users may substantially benefit from systems providing customized result abstracts for different tasks.

5.3.3 Clicks on Previously Visited Results

Similar to the default settings of web search, we show visited and unvisited URLs in different colors (purple and blue) in the experimental search system. Therefore, the users could quickly distinguish visited URLs from unvisited ones by color. We found that about 20% of the total fixations were on previously visited result abstracts (Table 3 "% fixations on visited") and there were 30%–40% chances that users will revisit a clicked result (Table 5 "P(click | examine, visited)"). This suggests that users still paid certain attention to the visited result abstract in SERP browsing rather than completely ignoring them, indicating that users may still expect to use the visited results when necessary.

However, results show that the chance of clicking an examined result is indeed lower than normal if the result has been previously visited by the users within the same session. In all types of tasks, the probability of clicking an examined result reduces if the result is previously visited by the users (comparing "P(click|examine)" and "P(click|examine,visited)"). However, the changes are more significant in the KS and IN tasks (decreased by 28% and 20%) compared to KI and EX tasks (by 9% and 8%). This suggests that whether the result has been visited or not has greater effects on users' clicking decisions in KS and IN tasks. Among four types of tasks, users in EX sessions are slightly more willing to re-open visited results (44%) comparing to other tasks (33%–39%), probably due to the complex nature of exploratory search tasks.

This section provides suggestions on how to deal with previously visited results in a search session. It seems risky to completely remove them because there are substantial needs to re-open previously visited results. However, the reduced chance of clicking suggests we may demote the ranking of the visited results in a new SERP. Besides, the results also show that we can customize systems for different tasks – e.g., for EX search, we may demote the rank of previously visited results less.

5.3.4 Clicks on Results at Different Ranks

Finally, we show click rates of results at rank R1–R9 in Figure 2. As before, we separately examine the click rates in the first view of each SERP and those in all SERP views. The click rate decays with the increase of result rank, but more quickly than the drop of fixation rate on result ranks shown in Figure 1.

As shown in Figure 2 (left), among the four tasks, we found that users in KS tasks have observably higher chance (about 10% more that on other tasks) of clicking the top one result but apparently lower chance of clicking the second top result in the first SERP

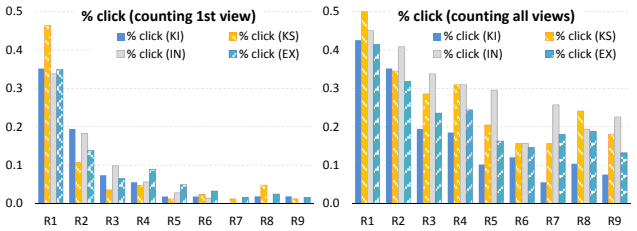

Figure 2. Click rates on results at each rank (R1 – R9), counting the first view (left) or all views (right) of each query.

view of a query. The reason is unclear, but this results in overall higher rates of clicking R1 in KS sessions compared with other tasks (as shown in the right figure). Similar to Figure 1, we also found that the chances of clicking results at lower ranks are significantly lower in KI tasks (counting all SERP views). In fact, users in KI tasks have the lowest chance to click R3–R9 among the four types of tasks, showing that in KI tasks users mainly focus their attention on the top ranked few results. The average and deepest rank of the clicked results in Table 7 also support this finding. This again suggests that we may tailor search systems by the types of tasks, e.g., generate best top few results in KI tasks.

5.4 Query Reformulation

Finally, we compare the four types of tasks by the way users issue and reformulate search queries, which indicates how users proceed through their search session.

5.4.1 Characteristics of Queries

Table 6 shows statistics of user queries and user behaviors for query reformulation. We notice that in different tasks, user queries vary in length, search effectiveness, and novelty.

As shown in the table ("# terms"), users issued significantly shorter queries (3.54 and 3.55 words) in tasks looking for factual information (i.e., KI and KS) comparing to those with intellectual search goals such as IN and EX tasks (4.39 and 4.38 words). The queries of the four types of tasks also vary in terms of effectiveness of retrieving relevant information. We calculate P@10, nDCG@10, and Reciprocal-rank for queries of different tasks and report the mean values in Table 7: users issued queries with better search effectiveness in KS and IN tasks.

However, it should be noted that the effectiveness of queries in IN tasks may be over-estimated. We further analyze the number of common results in multiple queries of the same session. For each query except the initial one of a session, we calculate the number of results retrieved by both this query and the previous query ("# overlap results") and Jaccard similarity between this query and its previous query's first page of results ("Jaccard similarity"). We can see that queries in IN sessions have significantly greater overlap of results (2.61 in common and 0.23 Jaccard similarity) than other tasks (0.81–1.33 results in common and 0.07–0.1 Jaccard similarity). Therefore, it is unclear whether queries in IN tasks are truly more effective because users may not be interested in some of the previously visited relevant results.

Table 6. Average user behavior statistics for a search query.

Statistics	KI	KS	IN	EX	
# terms	3.54	3.55	**4.39**	4.38	KI<IN**, KI<EX**, KS<IN**, KS<EX**
# overlap results	1.15	0.81	**2.61**	1.33	KI<IN**, KS<IN**, KS<EX*, IN>EX**
Jaccard similarity	0.09	0.07	**0.23**	0.10	KI<IN**, KS<IN**, IN>EX**
P@10	0.34	0.45	**0.47**	0.28	KI<KS**, KI<IN**, KI>EX*, KS>EX**, IN>EX**
nDCG@10	0.33	0.42	**0.48**	0.33	KI<KS*, KI<IN**, KS>EX**, IN>EX**
Recip-rank	0.70	0.76	**0.79**	0.68	KI<IN*, IN>EX*
Fix time results (transit)	1.06	0.93	**1.83**	1.39	KI<IN*, KS<IN*
Fix time results (normal)	**1.94**	1.35	1.46	1.91	KI>KS**, KI>IN**, KS<EX**, IN<EX**
Fix time task (transit)	**1.67**	1.52	0.87	0.82	KI>IN*, KI>EX**, KS>IN*, KS>EX*
Fix time task (normal)	**0.75**	0.39	0.29	0.11	KI>KS**, KI>IN**, KI>EX**, KS>EX**, IN>EX*
# use qsug / session	0.40	0.45	0.50	**0.60**	
Fix time qsug (transit)	0.33	0.75	0.62	**0.88**	KI<KS*, KI<EX*
Fix time qsug (normal)	0.03	0.06	0.05	**0.06**	

*, **: difference is significant at 0.05 and 0.01 level.

5.4.2 Source of Knowledge for Query Reformulation

Where do users acquire the knowledge for formulating new queries? To study this question, we look into user attention within the SERP views where users reformulated a search query (referred to as "transit SERP views"). We assume that if a user's attention on an area of the transit SERP view is apparently higher than those of a normal SERP view, they probably acquired knowledge from that area for query reformulation.

In Table 6, we label statistics in a transit SERP view by "(transit)" and those in a normal SERP view by "(normal)". For all the tasks, we observed increased attention of users on the task description and query suggestion in a transit SERP view. Additionally, users spent substantial time examining result abstracts in a transit SERP view. This indicates that task information, query suggestion, and result abstracts are possible sources of knowledge for users' query reformulation. Note that users do not necessarily need to adopt a query suggestion to be helped by one: they can get useful terms from the suggested queries (as suggested by Kelly et al. [18]). Results show that users examined diverse areas of the transit SERP in different tasks, indicating distinct source of knowledge for reformulation in different tasks.

We found that in tasks with factual goals (KI and KS), users rely mostly on task information itself for reformulating queries. As shown in Table 7, users in KI and KS tasks spent 1.67s and 1.52s in total examining task description in a transit SERP ("Fix time task info (transit)"), while in a normal SERP they spent only 0.75s and 0.39s. In addition, we noticed that users spent twice as much of the time on task description in tasks with factual goals compared with those with intellectual goals during a transit SERP view (0.87s and 0.82s). Also, in KI and KS tasks, the attention users put on task descriptions surpasses that on other areas of the transit SERP, such as the result abstracts (1.06s) and query suggestions (0.33s). All these results indicate that users in KI and KS sessions mainly focus on the task itself in query reformulation.

In comparison, we noticed that users in IN sessions probably reformulated queries mostly based on what they learned from the result abstracts. In KI, KS, and IN sessions, the total fixation duration on the result abstracts is shorter in a transit SERP view than those in a normal SERP view. However, in IN sessions, there is increased attention on result abstracts when reformulating queries (from 1.46s to 1.83s). Further, the amount of time users spent on examining result abstracts (1.83s) is also longer than they spent on task description (0.87s) and query suggestions (0.62s) in a transit SERP view.

Users in EX tasks are distinguished by the longest fixation duration on query suggestions in a transit SERP view (0.88s) among the four types of tasks. We also found increased attention of EX task users on task description during a transit SERP view (0.82s) compared with those in a normal SERP view (0.11s), indicating that task information may still constitute an important source of knowledge in EX tasks for query reformulation.

5.4.3 Use of Query Suggestion

Throughout the whole session, we found that users have limited direct use of query suggestion (i.e., adopting a query suggestion for search). As shown in Table 7, the number of times a query suggestion was used for search ranges from 0.4 to 0.6 in a session. Although we observed relatively higher frequencies of using query suggestion in exploratory search tasks, the differences are not significant and the usage frequency is still low (0.6). This may indicate the limited support of query suggestion for long search sessions in existing search engines.

To conclude, results suggest distinct strategies of supporting query reformulation in different tasks. For example, as users focus a lot on result abstracts in IN tasks, it may be preferable to generate suggestions based on frequent terms in result abstracts.

6. CHANGE OF BEHAVIOR IN A SESSION

How do search behaviors of users change in a search session? To answer this question, we compare users' search behavior in the initial query of a session with that in subsequent query reformulations. It should be noted that in our experiment we set a 10-minute limit for task completion. Therefore, the last query of each session was usually interrupted, and the behavior statistics may be inaccurate (e.g., with less SERP views, examined results and clicks). We did not consider this issue in the previous section because it does not introduce bias when we compare the differences between tasks. However, when comparing different queries in a search session, the last query of a session should be removed. Therefore, in this section, we selected 48 sessions with at least 3 queries and compare behaviors in the initial query with those in subsequent query reformulations excluding the last query of the session. The 48 sessions include 13 KI sessions, 9 KS sessions, 11 IN sessions and 15 EX sessions. Due to the limited sample size, we report significance at 0.1 level when necessary.

6.1 Decreased Interests on Search Results

We noticed that as a search session progresses, search results apparently attract less of the user's interest. As shown in Table 7, the number of unique clicks per query ("# uniq click / q") dropped significantly by 40%–60% in all tasks. The number of unique fixations per query ("# uniq fix / q") also decreased significantly, though by a smaller magnitude (about 20%–30%), except in IN sessions. Also, the number of SERP views per query ("# SERP views / q") reduced significantly by about 1–2 views per query. These all indicate that users became less and less interested in the results after a few searches.

We hypothesize three possible reasons for the reduced interests of users in a search session: 1) less relevant results are retrieved in subsequent query reformulations comparing to the initial query; 2) although as many as relevant results are returned, users lose their interests to the results because they are either highly overlap with results of previous queries or include very similar information; 3) users are becoming less persistent in SERP browsing.

We verify the first reason by comparing search effectiveness of query reformulations with those of the initial query in a session. We found that downgraded search performance may be one of the major reasons for KI and EX tasks resulting in decreased interests of users on results. Table 7 shows that the search performance of queries in KI and EX sessions are indeed decreasing, but there is no significant change of search effectiveness in KS and IN tasks. Both Reciprocal-rank and nDCG@10 declined significantly in KI and EX sessions. Due to the reduced search performance, it is not surprising that users may quickly feel that search results are not worth clicking and it is unnecessary to continue browsing a SERP after just one or two SERP views and clicks.

Table 7. Changes of search behaviors in query reformulations (excluding the last query) compared with the initial query.

Statistics	KI		KS		IN		EX	
# SERP views / q	2.92	1.70 ↓	3.78	1.89 ↓	3.45	2.80	3.93	1.80 ↓
# uniq fix / q	3.69	3.05 ▼	3.56	2.88 ▼	2.64	3.35	4.47	3.17 ↓
# uniq click /q	1.92	1.31 ▼	3.44	1.43 ↓	3.00	1.83 ↓	3.20	1.54 ↓
P(click | examine)	0.52	0.47	0.67	0.46 ↓	0.74	0.46 ↓↓	0.58	0.47
P(click | examine, rel)	0.64	0.64	0.95	0.62 ↓↓	0.77	0.57 ↓	0.70	0.65
Avg examine rank	2.66	3.19 ▲	4.20	3.75	3.66	3.79	4.05	3.74
Deepest examine rank	3.57	4.39 ▲	5.04	5.01	4.95	5.11	5.21	5.25
Avg click rank	3.34	3.43	4.50	3.88	4.09	3.95	4.40	4.04
Deepest click rank	3.43	3.43	4.50	3.88	4.09	3.96	4.45	4.05
Time view a SERP	14.9	9.7 ▼	12.7	7.9 ▼	8.7	11.4	16.6	8.7 ↓
Time a SERP view	4.95	5.82	3.35	4.15	2.51	4.51 ↑↑	4.23	4.90
Recip-rank	0.92	0.63 ↓	0.69	0.73	0.69	0.79	0.81	0.64 ▼
nDCG@10	0.46	0.26 ↓	0.40	0.32	0.46	0.41	0.37	0.29 ▼

▲/▼, ↑/↓, ↑↑/↓↓: difference is significant at 0.1, 0.05, and 0.01 level.

Further, we examine the validity of the second reason by the chances of clicking an examined result ("P(click | examine)") and an examined relevant result ("P(click | examine, rel)"). As shown in Table 7, as the session progresses, users in KS and IN sessions are less likely to click an examined result no matter whether it is relevant or not. For KS and IN sessions, we also did not find significant changes of queries' search performance in a session. Therefore, this indicates that it is probably the users themselves who believed that the search results, even the relevant results, are becoming less useful and worth clicking in a search session. One reasonable explanation could be that either the results are exactly previously retrieved ones, or similar information of the results appeared in previous results and users already knew relatively enough about it. Therefore, we conclude that declined novelty of search results may be one of the reasons resulting in decayed interest of users on search results in KS and IN tasks.

Finally, we examine whether users become less persistent in SERP browsing in a search session. Results indicate that users' persistence of browsing probably decreased in the tasks with unclear goals (KS and EX), but no evidence supports that users become less persistent in tasks with specific goals (KI and IN). As shown in Table 7, we found that the examined results in KS and EX tasks moved to higher ranked positions. The average rank of the examined results ("Avg examine rank") decreased from 4.20 to 3.75 in KS tasks and from 4.05 to 3.74 in EX tasks. Figure 5 also shows that, in KS and EX tasks, the chance of examining results decreased on every rank position without any exception. These all indicates that users in KS and EX tasks become less persistent and are more likely to stop browsing a SERP earlier than they did at the beginning of a search session. In comparison, the examined results in KI and IN tasks moved to lower ranked positions (the difference is significant in KI tasks). Also, Figure 5 shows that there are increased fixation rates on the results at lower ranked positions in the SERP. None of the evidences support decreased persistence of users in KI and IN sessions.

The decreased interests of users on search results indicate that users encountered difficulties as the search session progresses, but existing search systems did not provide supports for long sessions. It also partly confirms a hypothesis in search session performance evaluation that more weights should be put on the relevant results found at the early stage of a session [14]. Our studies of the three reasons also suggest different ways of supporting search sessions. For KI and EX sessions, the strategy is straightforward, i.e., it may help simply by improving search performance of queries. For KS and IN sessions, however, it requires systems that can retrieve novel search results without downgraded performance. For tasks with unclear search goals (KS and EX), we can optimize results for precision at higher ranked positions because users are less persistent to read lower ranked results of a SERP.

6.2 Changes of Browsing and Clicking

Figure 3–6 shows changes of fixation and click rates in the four types of tasks, counting the first view or all SERP views. Results show different changes of browsing patterns in the four tasks.

Figure 3 shows the changes of fixation rates in initial query and query reformulations, counting only the first view of each SERP. We noticed that throughout a search session, users shifted their attentions to focus less on the top 1 or 2 results but more on lower ranked results such as R3–R5. For example, in KI sessions, the chances of examining R4 and R5 increased, with less fixations on R1 to R3. Similarly, users moved their attentions from R1–R2 to

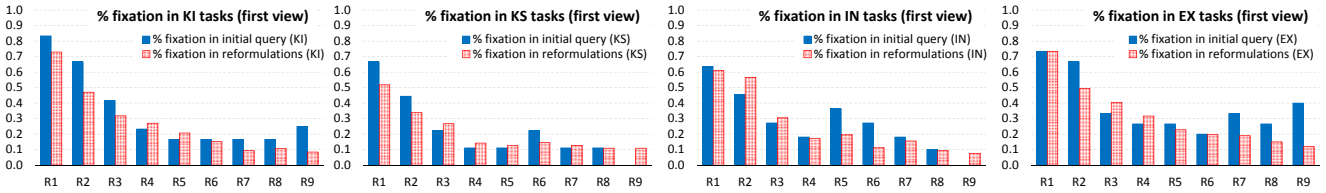

Figure 3. Changes of fixation rates in different tasks (initial query vs. query reformulations, counting 1st view of each SERP).

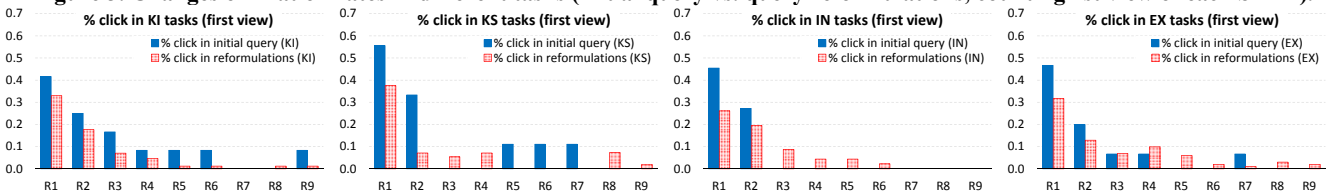

Figure 4. Changes of click rates in different tasks (initial query vs. query reformulations, counting 1st view of each SERP).

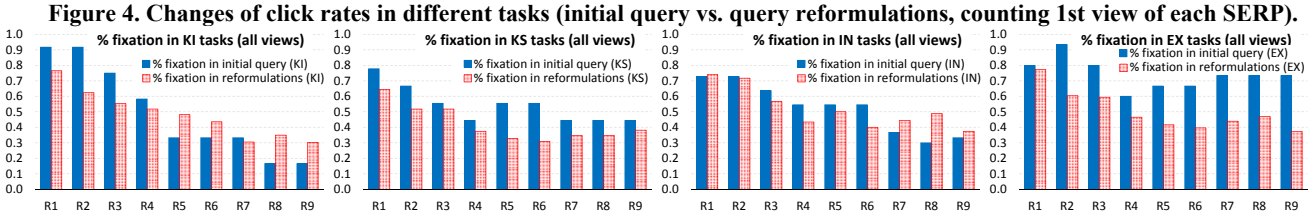

Figure 5. Changes of fixation rates in different tasks (initial query vs. query reformulations, counting all views of each SERP).

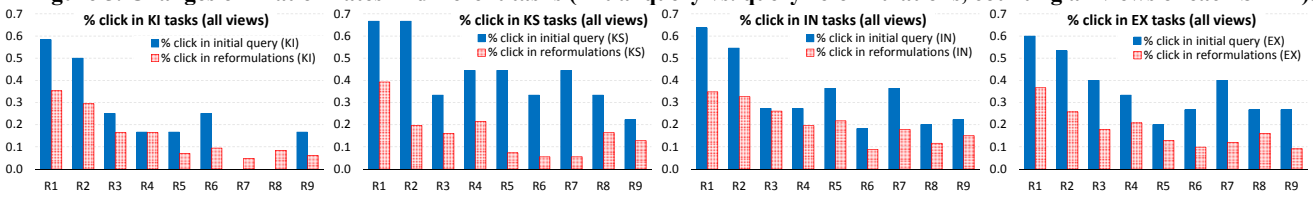

Figure 6. Changes of click rates in different tasks (initial query vs. query reformulations, counting all views of each SERP).

R3–R5 in KS tasks, from R1 to R2–R3 in IN sessions, and from R1–R2 to R3–R4 in EX sessions. However, users still mainly focused on the top results than others.

Figure 5 further shows the changes of fixation rates counting all SERP views of a query. As we discussed in the last section, the chance of examining a result decreased at every position in KS and EX tasks, but users moved their attentions from the top half of the SERP to the bottom results in KI and IN tasks. In addition, we note that there are some overall changes of browsing patterns in Figure 5. For KI and IN sessions, the slope of decreasing fixation rate by result rank is steep at the initial query of a session but less apparent in further query reformulations (this is due to decreased fixations on the top ranked results and increased attentions on the results at the bottom). In contrast, in KS and EX sessions, users' attentions are increasingly biased to the top ranked results. This is probably related to whether the goal is specific or amorphous.

As shown in Figure 4 and 6, the chances of clicking dropped significantly in almost all positions in four types of tasks, supporting our findings in the previous section. Though the chances of examining the top one result, as shown in Figure 3 and 5, did not drop by a large magnitude, the chances of clicking the top one result in query reformulations decreased to only about 2/3 to 1/2 of chances in the initial query.

The changing of browsing and clicking patterns indicate that, even during the session of the same tasks, we should customize the systems to support users at different time point of the search session. For example, due to the shifted pattern of fixation, in KI and IN sessions, systems may need to optimize search results for precision at very top positions at the beginning of a search session but shift to consider more on results at lower ranked positions after a few searches.

7. CONCLUSION

In this paper, we studied users' search behavior in long sessions of four different types of complex tasks. We found that search behavior varies distinctly by task and changes significantly after time. Table 8 summarizes our findings by four tasks and two dimensions.

Although it is confirmed that users' search behaviors will vary in different tasks, it is unexpected that only a small part of the differences show connections with the two task dimensions. In some cases, one type of task shows unique characteristics that are different from the other three. Sometimes we observed similarity between tasks that are different in both dimensions (e.g., KI and EX tasks, and KS and IN tasks). In addition, some characteristics exist in all types of tasks. This indicates that the two dimensions (product and goal) are probably still not enough to fully explain the differences of tasks and the underlying mechanisms that make user behavior different. Currently it remains unclear what the other possible factors are and how they might be identified.

One unique contribution of our work is that the results provide suggestions for systems to support sessions of complex search tasks. Specifically, our analysis of browsing and clicking patterns on the basis of eye-tracking data suggests that systems should be tailored for the search task at hand and the specific time point in the session. This advocates for futures systems that can automatically detect types of search tasks and optimize systems for the corresponding tasks, with dedicated supports during the search session. It also challenges existing evaluation metrics with fixed parameters in browsing and clicking models [4, 25] during a search session [14, 16].

ACKNOWLEDGEMENTS

Table 8. Summary of findings by tasks and dimensions.

Known Item (KI) • More searches, fewer examined and clicked results (5.1) • Greater efforts examining SERPs (5.2.1) • Most biased to top results in browsing (5.2.2) and clicking (5.3.4) • Click selectively (5.3.1 & 5.3.2) • Less fixations per query (6.1) • Downgraded query search performance (6.1)	**Known Subject (KS)** • Fewer searches, more examined and clicked results (5.1) • More time reading result webpages (5.2.1) • Better query search performance (5.4.1) • Less fixations per query (6.1) • Less willing to click examined results (6.1)	**Product: Factual** • Higher click accuracy (5.3.2) • Shorter queries (5.4.1) • Focus on task information for query reformulation (5.4.2)
Interpretive (IN) • Fewer searches, more examined and clicked results (5.1) • More time reading result webpages (5.2.1) • Least likely sequential browsing (5.2.3) • Better query search performance (5.4.1) • Highest overlap of results (5.4.1) • Reformulate based on results (5.4.2) • Less willing to click examined results (6.1)	**Exploratory (EX)** • More searches, fewer examined and clicked results (5.1) • Greater efforts examining SERPs (5.2.1) • Lowest click accuracy (5.3.2) • Willing to re-open visited results (5.3.3) • Less fixations per query (6.1) • Downgraded query search performance (6.1)	**Product: Intellectual** • Less likely sequential scanning (5.2.3)
Goal: Specific • Increased fixations on lower ranked results after time (6.2)	**Goal: Amorphous** • Wider breadth of a SERP view (5.2.3) • Less persistent in browsing (6.1) • Decrease of fixations (6.2)	**All Tasks** • Substantial attentions on lower ranked results (5.2.2) • Over 20% total fixations on visited results (5.2.4) • More clicks on examined relevant result (5.3.2) • Limited use of "related searches" (5.4.3) • Less SERP views and clicks per query (6.1)

This work was supported in part by the School of Information Sciences at the University of Pittsburgh and the Center for Intelligent Information Retrieval at the University of Massachusetts Amherst. Part of the work was done when the first author was at the University of Pittsburgh. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] Bates, M.J. 1989. The design of browsing and berrypicking techniques for the online search interface. Online Information Review. 13(5): 407-424.

[2] Broder, A. 2002. A taxonomy of web search. SIGIR Forum. 36(2).

[3] Buscher, G. et al. 2010. The good, the bad, and the random. In Proc. SIGIR'10: 42-49.

[4] Chapelle, O. et al. 2009. Expected reciprocal rank for graded relevance. In Proc. CIKM'09: 621-630.

[5] Clarke, C.L.A. et al. 2007. The influence of caption features on clickthrough patterns in web search. In Proc. SIGIR'07, 135-142.

[6] Cole, M.J. et al. 2010. Linking search tasks with low-level eye movement patterns. In Proceedings of the 28th Annual European Conference on Cognitive Ergonomics.

[7] Cole, M.J. et al. 2011. Task and user effects on reading patterns in information search. Interacting with Computers. 23(4): 346–362.

[8] Cutrell, E. and Guan, Z. 2007. What are you looking for? In Proc. CHI'07: 407-416.

[9] Dahlberg, J. 2010. Eye Tracking with Eye Glasses. Master Thesis, Umea University.

[10] Dumais, S.T. et al. 2010. Individual differences in gaze patterns for web search. In Proc. IIiX'10: 185-194.

[11] Granka, L.A. et al. 2004. Eye-tracking analysis of user behavior in WWW search. In Proc. SIGIR'04: 478-479.

[12] Guan, Z. et al. 2007. An eye tracking study of the effect of target rank on web search. In Proc. CHI'07: 417-420.

[13] Jansen, B.J. et al. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. Information Processing & Management. 36(2): 207-227.

[14] Järvelin, K. et al. 2008. Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions. In LNCS 4956: Proc. ECIR'08: 4-15.

[15] Joachims, T. et al. 2005. Accurately interpreting clickthrough data as implicit feedback. In Proc. SIGIR'05: 154-161.

[16] Kanoulas, E. et al. 2011. Evaluating multi-query sessions. In Proc. SIGIR'11: 1053-1062.

[17] Kanoulas, E. et al. 2012. Overview of the TREC 2012 Session Track. In Proc. TREC 2012.

[18] Kelly, D. et al. 2009. A comparison of query and term suggestion features for interactive searching. In Proc. SIGIR'09: 371-378.

[19] Li, Y. and Belkin, N.J. 2008. A faceted approach to conceptualizing tasks in information seeking. Information Processing & Management. 44(6): 1822–1837.

[20] Liu, J. et al. 2012. Exploring and predicting search task difficulty. In Proc. CIKM'12: 1313-1322.

[21] Liu, J. et al. 2010. Search behaviors in different task types. In Proc. JCDL'10: 69-78.

[22] Liu, J. and Belkin, N.J. 2010. Personalizing information retrieval for multi-session tasks. In Proc. SIGIR'10: 26-33.

[23] Lorigo, L. et al. 2006. The influence of task and gender on search and evaluation behavior using Google. Information Processing & Management. 42(4): 1123-1131.

[24] Moffat, A. et al. 2013. Users Versus Models: What Observation Tells Us About Effectiveness Metrics. In Proc. CIKM'13.

[25] Moffat, A. and Zobel, J. 2008. Rank-biased precision for measurement of retrieval effectiveness. ACM TOIS 27(1): 2:1–2:27.

[26] Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. Psychological bulletin. 124(3).

[27] Spink, A. et al. 2002. Multitasking information seeking and searching processes. Journal of the American Society for Information Science and Technology. 53(8): 639–652.

[28] Spink, A. et al. 2001. Searching the web: The public and their queries. Journal of the American Society for Information Science and Technology. 52(3): 226–234.

[29] Thomas, P. et al. 2013. What Users Do: The Eyes Have It. In Proc. AIRS'13: 416–427.

[30] White, R.W. and Roth, R.A. 2009. Exploratory search: Beyond the query-response paradigm. Synthesis Lectures on Information Concepts, Retrieval, and Services. 1(1): 1–98.

[31] Wu, W.-C. et al. 2012. Grannies, tanning beds, tattoos and NASCAR. In Proc. IIIX'12: 254-257.