How does Domain Expertise Affect User's Search Interaction and Outcome in Exploratory Search?

JIAXIN MAO, Tsinghua University
YIQUN LIU, Tsinghua University
NORIKO KANDO, National Institute of Informatics
MIN ZHANG, Tsinghua University
SHAOPING MA, Tsinghua University

People often conduct exploratory search to explore unfamiliar information space and learn new knowledge. While supporting the highly dynamic and interactive exploratory search is still challenging for the search system, we want to investigate which factors can make the exploratory search successful and satisfying from the user's perspective. Previous research suggests that domain experts have different search strategies and are more successful in finding domain-specific information, but how domain expertise level will influence user's interaction and search outcomes in exploratory search, especially in different knowledge domains, is still unclear. In this work, via a carefully designed user study that involves 30 participants, we investigate the influence of domain expertise levels on the interaction and outcome of exploratory search in three different domains: environment, medicine, and politics. We record participants' search behaviors, including their explicit feedbacks and eye fixation sequences, in a laboratory setting. With this dataset, we identify both domain-independent and domain-dependent effects on user behaviors and search outcomes. Our results extend existing research on the effect of domain expertise in search and suggest different strategies for exploiting domain expertise to support exploratory search in different knowledge domains.

CCS Concepts: • **Information systems → Users and interactive retrieval**; *Web search engines*; *Personalization*;

Additional Key Words and Phrases: Exploratory Search, Domain Expertise, User Behavior Analysis

ACM Reference format:
Jiaxin Mao, Yiqun Liu, Noriko Kando, Min Zhang, and Shaoping Ma. 2018. How does Domain Expertise Affect User's Search Interaction and Outcome in Exploratory Search?. *ACM Transactions on Information Systems* 1, 1, Article 1 (January 2018), 29 pages.
https://doi.org/0000001.0000001

1 INTRODUCTION

Modern search engines enable their users to efficiently access the massive amount of information on the Web. As a result, search users tend to use them to learn new skills and knowledge during their search processes. When users search to acquire knowledge, their information needs are

Correspondence concerning to this article should be address to Yiqun Liu, Department of Computer Science and Technology, Beijing National Research Center for Information Science and Technology, Tsinghua University. This work is supported by Natural Science Foundation of China (Grant No. 61622208, 61732008, 61472206), National Key Basic Research Program (2015CB358700), and JSPS KAKENHI (Grant No. JP16H01756).

1046-8188/2018/1-ART1 $15.00
https://doi.org/0000001.0000001

often multi-faceted, open-ended, and sometimes not clear at the beginning. Therefore, their search sessions usually span multiple queries and involve rich interactions. These information-seeking processes are described as *exploratory search* by White and Roth [38].

While search engines are good at solving simple fact-locating tasks, supporting exploratory search is still considered challenging. Users often have the feeling that they have to struggle to find the right information during the search processes and they are often frustrated after the searches [33]. One of the reasons that make supporting exploratory search harder is that the user plays a very important role in the interactive search process. They often need guidance in exploring unfamiliar information domains [18]. Therefore, besides retrieving relevant results according to the issued queries, the search system needs to provide guidance to help the user with different knowledge backgrounds. Making search engines more effective in supporting exploratory search requires an understanding of the search process from user's perspective. In particular, we want to understand what *user factors* can affect the *user behavior* and *search outcome* of exploratory search.

User's *domain expertise* may be one of such factors. The concept of domain expertise in IR covers two aspects: the *declarative* domain knowledge (i.e. the user's "*knowledge of the subject area that is the focus or topic of the search*" [39]) and the *procedural* domain-specific search strategies [5]. Previous research has shown that domain experts are more likely to be successful in Web search than novices [37]; users' domain knowledge levels influence their querying behaviors [12, 40] and search effectiveness [40]; and the domain experts have domain-specific search strategies in terms of site selection and goal sequencing [5]. The domain experts usually have more domain knowledge and domain-specific search strategies at the same time. because the declarative knowledge can affect the procedural knowledge over time through procedural learning [1], identifying and characterizing the effects of domain knowledge and domain-specific search strategies separately can be tricky and is beyond the scope of this paper. Therefore, in this study, we focus on investigating how the general domain expertise affects the user interaction and search outcome of exploratory searches, and address three research questions using the data collected in a carefully designed lab-based user study:

- **RQ1:** What is the relationship between domain expertise and search outcomes?
- **RQ2:** Which user behaviors, including how the user formulates queries, interacts with SERPs, and reads the landing pages, will be affected by domain expertise in exploratory search?
- **RQ3:** Are the influences of domain expertise consistent across different knowledge domains?

The research framework is shown in Figure 1. We controlled the independent variable, domain expertise level, by asking the participants from three different *User Domains* to complete tasks from all three *Tasks Domains.* Instead of measuring domain knowledge level in a continuous spectrum (e.g. [40]), a dichotomy for domain expertise level was assumed: the participants working with **IN**-domain tasks in their subject domains were regarded as expert users, and those working with **OUT**-domain tasks as non-expert users. In this way, the domain expertise effect will be analyzed by comparing the search behavior and results in the in-domain and out-of-domain search sessions. This experimental design is similar to that adopted by Bhavnani [4, 5] but has two major advantages: 1) While Bhavnani only focused on two domains: online shopping and healthcare, we included three knowledge domains, **E**nvironment, **M**edicine, and **P**olitics, in our study, which allows us to investigate whether the observed effect is associated with a specific knowledge domain or independent across all the studied domains. 2) The participants in different subject domains were undergraduate students hired from different departments of our university (see Section 3.1.2 for

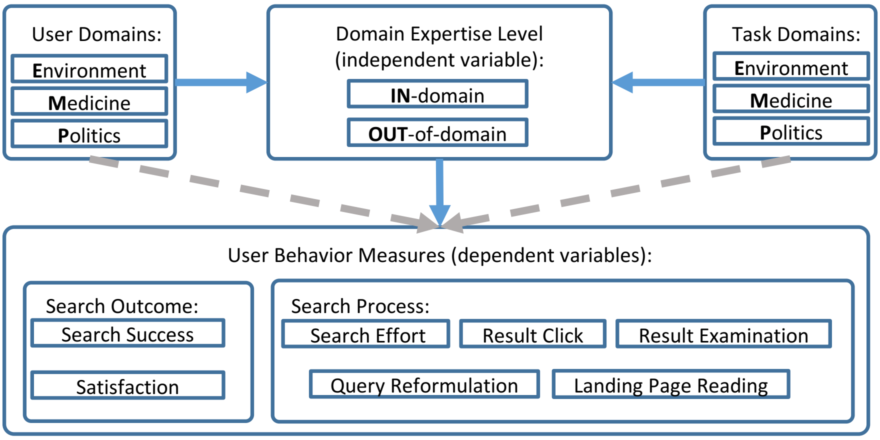


Fig. 1. Research Framework

details), which can be generalized to more knowledge domains easily if more experiment resources are given.

At the cost of a limitation of the experiment scale, the lab-based user study allows us to collect a comprehensive search behavior logs, which include participants' querying, clicking, mouse movement, tab-switching, and eye-fixation behaviors as well as participants' explicit feedback and questionnaire data. This rich dataset enables us to investigate the questions that are not covered in previous research. For example, with the help of eye-tracking devices we studied how domain expertise affects the origins of query terms.

The rest of the paper is organized as the following: Related studies are discussed in Section 2. The methodology used is described in Section 3. We present the results in Section 4. Finally, we discuss the findings in Section 5 and make conclusions in Section 6.

2 RELATED WORK

2.1 Exploratory Search

Marchionini [32] divided search activities into two broad categories: lookup and exploratory search, and associate exploratory search with learning and investigating activities. White and Roth [38] further characterized the exploratory search as "an information-seeking problem context that is open-ended, persistent, and multifaceted, and information-seeking processes that are opportunistic, iterative, and multi-tactical".

Due to its intrinsically complex and highly interactive nature, the exploratory search is challenging for both the users and the IR systems. Therefore, a lot of research aimed at understanding and supporting exploratory search. For example, to understand users' behavior in exploratory search, Jiang et al. [19] thoroughly studied users' browsing and clicking behavior for long search sessions solving complex search tasks. Athukorala et al. conducted user studies to estimate the subjective specificity of search results in exploratory search [3] and separate exploratory search from lookup search based on users' search behavior [2]. To better support exploratory search, Hassan Awadallah et al. [18] mined a large scale search log to provide users with task recommendations in complex tasks. Kong et al. [22] extended faceted search to general web search to assist exploratory search. In this study, we extend exploratory search understanding by investigating domain expertise effects and their implications for improving support for exploratory search.

Exploratory search often involves knowledge acquisition and some recent studies have attempted to assess the learning outcomes and knowledge development during the search process. Egusa et al. [13] used a concept map method to evaluate the changes in users' knowledge structure after searching. Collins-Thompson et al. [11] explored indicators of learning in a lab-based user study. Eickhoff et al. [15] identified the evidence of within-session learning activity in query logs. To better measure the knowledge acquisition during exploratory search, users' prior knowledge about the search task and its effect on the process and outcome of the search task should be considered. The findings of this study may be useful in developing better measures for the learning process and outcomes of the learning-related exploratory search in the future.

2.2 Domain expertise in search

Domain expertise has been studied extensively in the IR community. We divide existing studies into two broad categories: users' domain expertise level is *measured* in the study; or the domain expertise level is explicitly *manipulated* by experimental settings.

In the first category, the researchers designed methods to measure users' domain expertise levels, especially their domain knowledge levels. Zhang et al. [40] accessed the domain knowledge level by having participants rate their familiarity with terms from a domain-specific thesaurus in the engineering and science field. Similar domain knowledge assessment method was applied to medicine and biology domains by Cole et al. [9] to study how domain knowledge level affects search users' low-level eye movement patterns and by Liu et al. [26] to investigate how experts and novices adapt their query reformulation strategies according to the task difficulty. Zhang et al. [41] further exploited multiple regression models to predict the domain knowledge level measured in this way with observed behavior variables. Thesaurus-based methods have been widely adopted and have proven effective in assessing participants' domain knowledge levels. However, the requirement of domain-specific thesauri makes it hard to generalize the measurement across different domains.

Another common way to elicit users' domain knowledge level is by testing participants' answers to knowledge quiz questions. Duggan and Payne et al. [12] had participants answer 15 questions in both football and music domains and measured their domain expertise level in each domain with the number of correctly answered questions. Their study found that higher domain knowledge level can be associated with better search performance. Kang and Fu [20] combined both self-reported ratings and knowledge quiz questions to measure domain expertise level in finance and economic domains and compared how expert and novice users perform differently in using Web search engines and the social tagging system Delicious.

Users' domain expertise can also be inferred from their interaction history. White et al. [37] identified domain experts in a naturalistic log-based study through specific websites visited by users. They inspected the domain expertise effect in four different domains and built prediction models to estimate real users' knowledge level. We also try to examine the commonalities and differences of domain expertise effects across domains. Compared to a log-based study, our lab-based user study enables us to control the variability of search tasks and inspect a variety of data that could not be obtained in search logs such as eye-tracking sequences and users' explicit feedback.

In the second category of studies focusing on domain expertise, several methods were used to manipulate participants' expertise levels. Longitudinal user studies were conducted to investigate the changes of participants' search tactics while their domain knowledge evolves during a relatively long period of time [36, 39]. An issue with such extended experiments is that participants' expertise can change as well as search tasks, making it hard to isolate domain expertise effects on behaviors and outcomes. Tamine and Chouquet [35] conducted a crowdsourcing study, in which the expert participants and non-expert participants were recruited in different platforms, to investigate how domain expertise influences query reformulation and relevance assessment in clinical settings. Bhavnani [4, 5] manipulated expertise levels by hiring online shopping experts and healthcare experts and having them perform both in-domain and out-domain search tasks. He found that domain expertise influenced users' site selection and goal sequencing patterns. Because the online shopping related tasks and healthcare related tasks are intrinsically different (e.g. the stopping criteria for shopping tasks is to find the lowest price) and they may not be representative cases for Web search, the identified domain-specific search strategies may be limited to these two domains and the specific tasks.

Since most of the existing research on domain expertise effect in search focused on one or two specific knowledge domains [35, 36, 41], whether the findings are consistent across different domains and generalizable to other domains has not been well investigated. White et al.'s large-scale log-based study [37] covered four different knowledge domains but because of the nature of log-based study yet could not: 1) fully control the search tasks; 2) analyze user behavior that can not be remotely logged, due to the nature of log-based study. To fill this research gap, we designed a lab-based user study that covered multiple knowledge domains and collected fine-grain user behavior logs, including participants' eye-fixation sequence and explicit feedback.

Our study differs from the exiting research in the following aspects: First, we proposed a new method to control participants' domain expertise level in lab-based user studies. The experiment settings are similar to those adopted by Bhavnani [4, 5]. However, compared to Bhavnani's studies that compared two very different search scenarios, all the search tasks or *simulated work tasks* [6] adopted in our study are answering an informational question using the information found during the search. We chose informational questions from three different knowledge domains and hired participants from the corresponding departments as domain expert users to control the domain expertise level. Compared to the log-based study conducted by White et al. [37], in our lab-based user study, we can fully control the experimental apparatus and search tasks completed by the participants, which enables us to collect richer measures and leads to different findings (see Section 3.5.1 for an example). Compared to the studies that use domain-specific thesaurus and quizzes to measure participants' domain knowledge level, the proposed method can be generalized to other knowledge domains more easily, and therefore, is more appropriate in identifying the effects of domain expertise levels instead of the effects associated with a specific domain.

Second, while it is impossible to cover all different domains in a single user study, we compared and analyzed the observed effects across these three representative domains. Compared to existing works, including three domains in the study enable us to inspect whether the effects of domain expertise are domain-independent or domain-specific. By investigating whether the effects are consistent across multiple knowledge domains, our study verifies the findings in some previous studies that focus on the domain expertise effects in a specific domain and provides further evidence for the domain-independent effects on users' search process and search outcome.

Third, we collected a rich dataset using a Tobii X2-30 eye-tracker and a dedicated browser plugin. With this fine-grain user behavior dataset, we investigated some open questions that were not covered in existing works. For example, by using the eye-tracker to capture all the text read by the participants during the experiment, we inspected how domain expertise affects the origins of novel query terms issued by the participants. The analysis of the origins of query terms directly tested the hypothesis that the domain expert users are better at issuing queries in exploratory search because they have richer domain-specific vocabulary and rely less on the terms encountered during search.

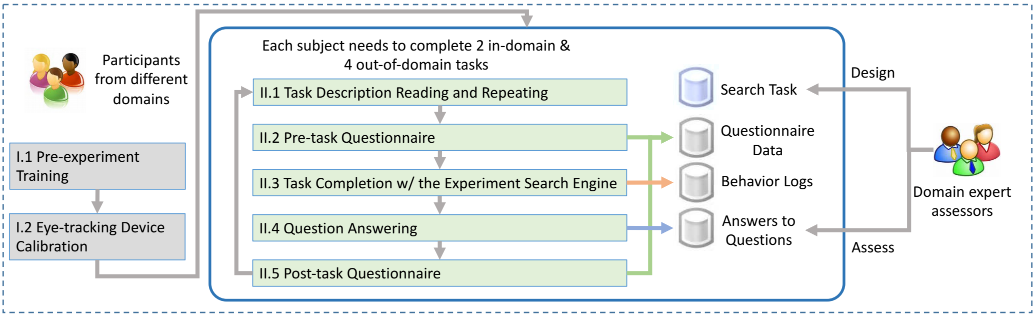

Participants from different domains
I.1 Pre-experiment Training
I.2 Eye-tracking Device Calibration
Each subject needs to complete 2 in-domain & 4 out-of-domain tasks
II.1 Task Description Reading and Repeating
II.2 Pre-task Questionnaire
II.3 Task Completion w/ the Experiment Search Engine
II.4 Question Answering
II.5 Post-task Questionnaire
Search Task
Questionnaire Data
Behavior Logs
Answers to Questions
Design
Assess
Domain expert assessors

Fig. 2. The user study procedure.

3 METHODS

3.1 User Study

We designed a user study to collect users' search behavior and feedbacks while controlling their domain expertise levels. In this section, we introduce the procedure (shown in Figure 2) of the user study in details. Note that the language used in the study is Chinese, so all the task descriptions, search systems, and instructions are in Chinese. We will show the translated text in the paper and attach the original text in the appendix.

3.1.1 Tasks. We selected 6 search tasks from three knowledge domains from our previous work [24] and recruited participants with corresponding background knowledge (See Section 3.1.2). The search tasks are demonstrated in Table 1.

The selected knowledge domains were environment, medicine, and politics. These three domains cover both social science and natural science areas. We intentionally avoided more popular domains like computer science and economics because we wanted to further ensured that a participant who was not from the corresponding domain has a relatively low domain knowledge level.

The search tasks were designed by three domain expert assessors who are graduate students in the corresponding subjects. Each search task is defined by an open-ended question that can be answered with 60-100 words. The "Product" aspect of each task listed in the last column. This aspect of search task was conceptualized by Li and Belkin [25] and later adopted in TREC 2012 Session Track [21]. The "Product" aspect varies between "Intellectual" and "Factual". While Factual tasks involve locating facts and data, Intellectual tasks are related to the production of new ideas or findings. Most of the tasks used in this study are categorized as "Intellectual" except for one multi-aspects "Factual" task in the medicine domain, therefore, the participant needs to issue multiple queries to accomplish each of them. Pilot experiment showed that these selected tasks have appropriate complexities and difficulties for the experiment. While they are not too difficult to complete for the novice participants, they are non-trivial and the answers to them have multiple aspects, therefore, even the domain expert participants needed to submit multiple queries to the search system to complete them.

We also asked the domain expert assessor to make ground truth answers and scoring criteria for the tasks designed by him or her as well as assign a score of 0-10 for every answer from the participants to measure the outcomes of the search (see Section 3.1.5 for more details). The ground truth answer usually contains 3-5 key points with different importance scores. The answer score can be computed by summing up the importance scores of key points covered by the submitted answer.

Table 1. The search tasks adopted in the user study.

Domain	Task ID	Product Category	Task Description
Environment	E_1	Intellectual	What are the characteristics of particle pollution (also called particulate matter) in China? Your answer should cover its compositions, its time-varying patterns, and its geographical characteristics.
	E_2	Intellectual	Why can't Ultraviolet (UV) disinfection completely supplant chlorination in disinfecting drinking water?
Medicine	M_1	Factual	What are the most commonly-used treatments for cancer in clinical? And what are the advantages and disadvantages of them?
	M_2	Intellectual	What are the potential applications of 3D printing for “Precision Medicine"?
Politics	P_1	Intellectual	Political scientist have noted that the trend of political polarization during the US presidential election is increasingly evident. What are the reasons behind it? (polarization here refers to the divergence of political attitudes to ideological extremes.)
	P_2	Intellectual	In order to achieve their own interests, what kind of strategies are often taken by US interest groups?

Table 2. The distribution of participants by college year

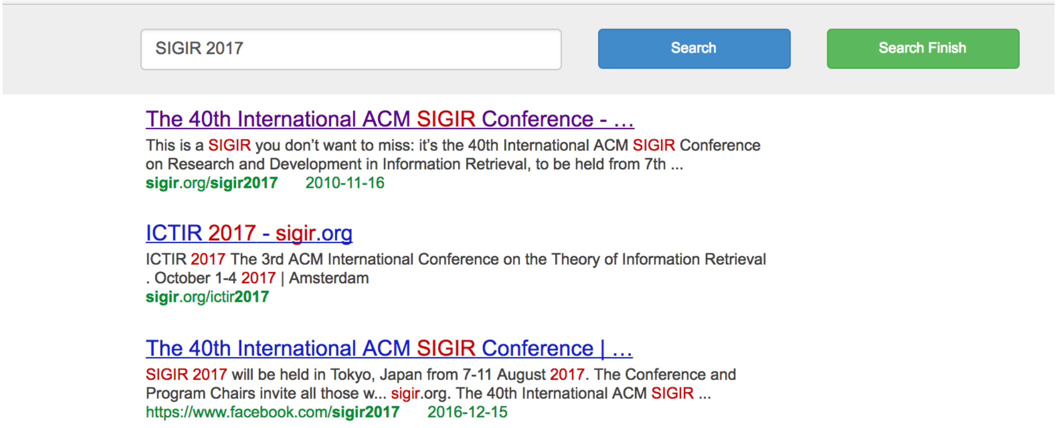
User Domain	2nd year	3rd year	4th year
Environment	2	3	5
Medicine	4	5	0
Politics	5	4	1
All	12	12	6

3.1.2 *Participants.*

As shown in Figure 1, we control the domain expertise level by hiring participants from corresponding user domains to complete both in-domain and out-domain search tasks.

For the environment, medicine, and politics domains, we sent recruiting message via email and online social networks to the environment, medical, and social science schools of Tsinghua university. A total of 30 participants took part in the user study, 10 from each target school. 22 participants were female and 8 were male. The age of participants ranged from 19 to 22. All the participants were native Chinese speakers and had college-level reading and writing skills. All the participants used Web search engines regularly for both studying and other daily purposes, so they were familiar with the search system and had an adequate level of general search expertise.

To ensure the participant has a reasonable domain knowledge level for the in-domain tasks, we did not allow first-year students to participate in the experiment. The distribution of college years of all participants are shown in Table 2. There were more 4th year students from the environment



SIGIR 2017

The 40th International ACM SIGIR Conference - ...
This is a SIGIR you don't want to miss: it's the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, to be held from 7th ...
sigir.org/sigir2017 2010-11-16

ICTIR 2017 - sigir.org
ICTIR 2017 The 3rd ACM International Conference on the Theory of Information Retrieval . October 1-4 2017 | Amsterdam
sigir.org/ictir2017

The 40th International ACM SIGIR Conference | ...
SIGIR 2017 will be held in Tokyo, Japan from 7-11 August 2017. The Conference and Program Chairs invite all those w... sigir.org. The 40th International ACM SIGIR ...
https://www.facebook.com/sigir2017 2016-12-15

Fig. 3. The screen-shot of the experimental search engine.

domain. We acknowledge this as a limitation of the user study, but we argue that it would not greatly affect the experiment results because we used the method introduced in Section 3.2 to control the effect from user domains. We also note that except the college year and the domain expertise level in each domain, the participants in our study have similar backgrounds and characteristics, which will reduce the undesired noise introduced by individual differences but may limit the generalizations of the experiment results to a broader user group.

Each participant completed all 6 search tasks, 2 in-domain tasks and 4 out-domain tasks. We assumed that the participant will have more background knowledge about the in-domain search tasks. Therefore, by designing search tasks from 3 task domains and hiring participants from 3 user domains, we managed to control the domain expertise level.

3.1.3 Search System. To simulate the Web search environment, we built an experimental search system that provides the participants with modified results from a commercial Web search engine. As shown in Figure 3, the experimental search engine has an interface similar to common Web search engines. A search finish button was added to allow the participant to stop the search and go to next stage. The search system supported query reformulation and pagination (up to first 5 pages). When the system receives a query, it will forward the request to Bing Search API to retrieve the first 50 organic search results. We filtered out all query suggestions, sponsored search results, and vertical results on the SERPs because existing research showed that these elements will affect user's examination and click patterns on the SERPs [29] and we want to focus on the influence of domain expertise on homogeneous results in this paper. The influence of domain expertise on heterogeneous results will be left to future research. We verified the response time of the search system is comparable to a commercial Web search engine in the pilot experiment. To avoid the potential effects caused by the personalized results, we also cached the organic results so that if other participants issued the same query, the same SERP would be returned.

3.1.4 Logging of Search Behavior. In the user study, we used an eye-tracking device and an extension for Chrome browser to unobtrusively log participants' search behaviors.

A Tobii X2-30 eye-tracker was deployed in the study to capture participants' eye movements on a 17' LCD screen with a resolution of 1366×768. For each participant, we used the standard 9-point calibration provided by Tobii Studio software before the experiment. To identify users' examination and reading behaviors, we adopted the I-VT filter from Tobii Studio to detect eye

Table 3. Some examples of task assignments in the user study.

Task Position:	1	2	3	4	5	6
assignment 1	E_1	M_2	P_1	E_2	M_1	P_2
assignment 2	M_1	E_1	P_2	M_2	E_2	P_1
assignment 3	P_1	M_2	E_2	P_2	M_1	E_1
assignment 4	E_2	P_2	M_2	E_1	P_1	M_1

fixations. By analyzing the fixation sequence, we can investigate how participants with different domain knowledge levels browse and examine SERPs as well as how they read landing pages.

Previous studies have used injected javascript to record users' clicking, scrolling, and mouse movement behaviors on SERPs [28, 31]. Because we also wanted to log participants' behaviors on the landing page, we implemented a Chrome extension that can inject javascript into every Web pages visited by the participants during their search processes. We further verified that the Chrome extension would not cause noticeable lagging during the experiment in the pilot experiment.

Besides the clicking, scrolling, tab-switching, and mouse-moving behaviors, the extension also: 1) recorded the sources as well as the bounding box coordinates and text content of every HTML element of the Web pages; 2) allowed the participant to give usefulness feedback on landing pages using right-click pop-up menu. With the bounding box and text content information and the fixation sequence recorded by the eye-tracker, we can capture the *text content* (on both SERPs and landing pages) that was actually attended to and read by the participant. We used a 4-level graded usefulness feedback (1: not useful at all; 2: somewhat useful; 3: fairly useful; 4: very useful) [31]. The default option is 1: not useful at all, so the participant only needed to mark useful landing pages encountered during search. We acknowledge that the participant could forget to make usefulness feedback in some search sessions. During the experiment, the experimenter would intervene and notify the participant to make usefulness feedback when the participant forgot to provide any usefulness feedback in a search task.

3.1.5 User Study Procedure. The procedure of the user study is shown in Figure 2. Before the experiment, we used an example task to demonstrate the procedure of the user study in the Pre-experiment Training stage (I.1) and calibrate the eye-tracker (I.2) for each participant. After that, the participant was asked to complete 6 search tasks, including 2 in-domain tasks and 4 out-domain tasks. To control for the potential order effect, the order of the tasks were assigned in the following way[1]. For each participant, we first generate a random permutation of three domains. Then, we assigned first 3 tasks by randomly selecting one of the two tasks in each domain sequentially, according to the permutation. Finally, we assigned the remaining 3 tasks according to the same permutation of domains. In this way, we ensured that each task has an equal probability to be assigned in one of six positions and two tasks from the same domain would not be assigned to two consecutive positions. Some examples of the task assignments are shown in Table 3.

To complete each task, the participant was required to go through 5 different stages (II.1-II.5). First, in II.1 stage, to make sure the participant can remember the search task during searching, he or she was instructed to read and memorize the task description in a Web page, and then re-input

[1] We tested the order effects after the user study by correlating some search outcome measures (satisfaction and answer score in Table 9) and search effort measures (#queries and task time in Table 10) with the task position (1-6). While the task position has no significant effects on satisfaction, answer score, and #queries (with all $p > 0.05$), task time is negatively correlated with task position (Pearson's $r = -0.174$, $p = 0.025$). This result confirms that the random assignment of search tasks was necessary and it can be explained by: 1) that the participants got familiar with the experimental scenario as they proceeded; and 2) that the participants got tired so they chose to use less time to complete the later tasks.

Table 4. The questions in the pre-task questionnaire (II.2 in Figure 2)

Measure	Question
Pre-knowledge	How much do you know about the topic of the task?
Pre-difficulty	How difficult do you think it will be to complete this search task?
Pre-interest	How interested are you to learn more about the topic of this task?

Table 5. The questions in the post-task questionnaire (II.5 in Figure 2)

Measure	Question
Post-knowledge	How much did your knowledge increase as you searched?
Post-difficulty	How difficult was this task?
Post-interest	How much did your interest in the task increase as you searched?
Satisfaction	How satisfied were you with your search experience?

the task description without viewing it on the next page. Then, we used a pre-task questionnaire to assess the participant's self-reported domain knowledge level (pre-knowledge), expected difficulty (pre-difficulty), and interest level (pre-interest) of the current task (II.2). The questions in the pre-task questionnaire are shown in Table 4. The participant was required to answer these questions using a 5-point Likert scale (1: not at all, 2: slightly, 3: somewhat, 4: moderately, 5: very). The results collected in the pre-task questionnaire can test whether our experiment design effectively manipulates the domain expertise levels.

After completing the pre-task questionnaire, the participant would be directed to the experimental search engine and would start performing the search tasks (II.3). The participants could issue queries, click on the results, and acquire information relevant to the task freely, just like using a normal Web search engine in daily life except they were required to use the right-click popup menu to mark landing pages that were useful for the task. While no hard time limits were imposed, we explicitly told the participant the typical searching time for each task is about 10-15 minutes. The participant could click the "search finish" button and stop searching when he or she thought the acquired knowledge was enough for answering the question, or no more useful information would be found.

After searching, the participant would answer the task-related question (II.4) and complete a post-task questionnaire (II.5). We asked the domain expert assessor who designed the search task to make a reliable *first-tier assessment* of the quality and correctness of the answers using a scale range from 0–10. The questions of the post-task questionnaire are listed in Table 5. We used the scores of the answers and the results collected in the post-task questionnaire to derive objective and subjective measures for the search outcomes.

The total time span for the experiment is about 2 hours and we paid the participants at the rate of about $8 per hour. To the participant to perform the task more thoroughly, we rewarded the participants who have top 5 highest average answer scores with an extra 40% of the payment.

3.1.6 Collected Data. Through the user study introduced in this section, we collected a search behavior dataset that contains 180 search sessions from 30 participants on 6 search tasks from 3 different domains. After an inspection of the dataset, we found that the eye-tracker malfunctioned in some of the search sessions. So we filtered out the problematic sessions. The number of participants and the number of search sessions in the remaining dataset are shown in Table 6. On average, the participants issued 3.93 queries, clicked 7.19 documents, and spent 544 seconds on searching for

Table 6. Statistics of valid search behavior logs

	#participants	#search sessions
Environment	10	58
Medicine	9	54
Politics	9	54
All domains	28	166

each task. The average length of issued queries was 9.93 Chinese characters, which indicates the issued queries in our study are relatively complex.

3.2 Data Analysis Method

To inspect the influence of domain expertise on participants' search outcome and search process, we use the domain expertise level, operationalized by whether the task is an in-domain task or out-domain task for the participant, as the independent variable and investigate its relationship with over a variety of dependent variables that comprehensively measure participants' search outcome and process (see Section 3.4 and 3.5 for the detailed descriptions of the measures).

Specifically, we divide the search sessions into two groups: 1) **IN**-domain sessions in which the participants performed search tasks from their own domain of expertise; 2) **OUT**-domain session in which the user domain and task domain do not match. We investigate the relationship between the domain expertise levels and user behavior measures, such as the answer score for the search outcome and the number of queries for the search process, by comparing the values of those measures in these two groups.

Because the independent variable (i.e. the domain expertise level) is controlled by the domain of participants and the domain of tasks, its relationships with the dependent variables are inevitably affected by the other effects associated with search tasks and user domains. For example, because of the nature of the tasks in politics domain, the participants from all three domains issued fewer queries in the politics domain than in other domains. If we do not take the effects caused by the tasks into consideration, we may discover an artifact that the domain expert users from the politics domain issue fewer queries in the in-domain sessions. Therefore, in this study, we use the following method to control the task and domain effects when exploring the potential effects of the domain expertise on the dependent variables.

First, for the effects caused by the tasks, we compute the relative deviation d_{ij} for each dependent variable of each search session:

$$d_{ij} = \frac{y_{ij} - \overline{y}_{\cdot j}}{\overline{y}_{\cdot j}} \tag{1}$$

Here y_{ij} is an original dependent variable corresponding to the search session conducted by participant i when completing search task j and $\overline{y}_{\cdot j}$ is the mean of all y_{ij} associated with task j. The relative deviation d_{ij} can be interpreted intuitively. For example, if whole user group on average spent 10 minutes to complete a search task j while a specific user i spent 5 minutes, the corresponding $\overline{y}_{\cdot j}$ and y_{ij} will be 10 *mins* and 5 *mins*. Then, d_{ij} is given by $\frac{5\ mins - 10\ mins}{10\ mins} = -0.5$, which indicates that user i spent 50% less time than average users.

Second, for the effect caused by the user domains, we separately report the average relative deviation (average d_{ij}) for both the **IN**-domain search sessions and **OUT**-domain search sessions that were generated by the participants from three different user domains. If the independent variable, domain expertise level, has an effect on the dependent variable, the d_{ij} of the in-domain

Table 7. The results of post hoc statistical power analysis. We compute the statistical power $(1-\beta)$ of Mann-Whitney U test given the designed sample sizes of the study, a significant level $\alpha = 0.05$, and different population effect size parameters measured in Cohen's d. ds of 0.2, 0.5, and 0.8 are defined as small, medium, and large effects respectively by Cohen [8].

	Designed Sample Size		Population Effect Size Parameter		
	$n_{\text{in-domain}}$	$n_{\text{out-domain}}$	small ($d = 0.2$)	medium ($d = 0.5$)	large ($d = 0.8$)
domain-specific effect	~20	~40	0.11	0.42	0.80
domain-independent effect	54	112	0.22	0.83	0.99

sessions should be different with those of the out-domain sessions. We use Mann-Whitney U test (a nonparametric alternative to the independent t-test [30]) for statistical significance because most of the dependent variables have a non-normal distribution.

Although we control the domain expertise level by the interaction between two factors, user domains and task domains, we did not use a conventional two-way ANOVA. First, two-way ANOVA can avoid the influence from user domain and task domain but not the influence of search task, given that there are two search tasks in each task domain. Also, the interaction effects revealed by two-way ANOVA are not necessarily caused by the domain expertise level. We would have to use additional methods to compare the measures in the in-domain and out-domain sessions; Finally, most of the dependent variables in the study are not normally distributed.

Note that the relative deviations of in-domain sessions and out-domain sessions from a single user domain can be both positive or both negative, which indicates the user domain has an effect on the dependent variable. We can still investigate the effect of domain expertise on this dependent variable by comparing these two relative deviations.

Finally, for the dependent variables with positive significant test results, we show the average relative deviation (with its standard errors) under different user domains and tasks domains jointly in figures (see Figure 4 for an example). From these figures, we can inspect whether the corresponding effect of domain expertise is domain-independent. If the difference between in-domain and out-domain session is *statistically significant* on the whole dataset, and it is *consistent* across all three user domains, then we are confident that the difference is caused by the effect of domain expertise level but not associated with specific search tasks or knowledge domains. If the difference is only significant for a specific domain and it is not consistent in other domains, then the corresponding effect is more likely to be domain-specific.

A post hoc statistical power analysis [8] was conducted using the `G*Power 3` program [16]. Setting the significant level $\alpha = 0.05$, given the sample sizes (about 20 in-domain sessions and 40 out-domain sessions per user domain), the powers $(1-\beta)$ of Mann-Whitney U tests with different population effect size parameters (measured by Cohen's d) are shown in Table 7. The conventional threshold of statistical power $(1-\beta)$ for a fair chance to reject incorrect null hypotheses is 0.8. Therefore, the results in Table 7 indicate that our study can effectively detect both medium and large domain-independent effects as well as large domain-specific effects.

3.3 Validating the Experiment Settings with Pre-task Questionnaire

Before investigating the effects of domain expertise on other measures, we want to test whether the experiment settings can effectively control participants' domain expertise levels in different search tasks. We collected the following feedbacks in the pre-task questionnaire stage (II.2 in Figure 2):

- **Pre-knowledge:** the participant's self-reported level of domain knowledge ;
- **Pre-difficulty:** the participant's expected task difficulty level;
- **Pre-interest:** the participant's task interest level.

If the experiment setting is valid, the expert user who is going to complete an in-domain search task will report a higher domain knowledge level than the non-expert user facing an out-domain task. We also expect to see the expert user reporting a higher interest level and a lower expected difficulty level for the in-domain search task, which will further validate that the experiment setting indeed has an effect on participants' domain expertise level.

3.4 Measures for Search Outcomes

To address **RQ1**, we define some measures to quantify search outcomes. In this study, we consider two aspects of search outcomes: search *success* and search *satisfaction*.

Because the search task is to answer informational questions, search success can be defined as successfully acquiring relevant and useful knowledge during search and correctly answering the question after search. Therefore, we use two measures for search success:

- **Post-knowledge**: the participant's self-reported knowledge gain in the post-task questionnaire;
- **Answer score**: the score of the submitted answer assessed by the domain expert assessors.

While the self-reported knowledge gain is a subjective measure of knowledge gain during search, the answer score is a rather objective measure of whether the participant found the correct information for the search task.

Search satisfaction is mainly associated with users' subjective feelings about their interaction with the search system and the whole search process. Therefore, we rely on the post-task questionnaire to quantify participants' satisfaction through the following measures:

- **Post-difficulty:** the participant's perceived difficulty level of the search task after search;
- **Post-interest:** the increase of the participant's interest level of the search task after search;
- **Satisfaction:** the self-reported search satisfaction.

We hypothesize that because the existing domain knowledge can help the user in information seeking, domain expert users are more likely to be successful in in-domain search tasks, and therefore, in general are more satisfied with the search process. If this hypothesis holds, domain expert users will: 1) have higher knowledge gain and answer scores when completing an in-domain tasks; and 2) report a higher level of search satisfaction along with a lower level of perceived difficulty and a higher level of interest increase in the post-task questionnaire.

3.5 Measures for Search Process

For **RQ2**, we inspect a variety of dependent variables which measures various aspects of the search process, including user search effort, query reformulation strategies, SERP examination and clicking, as well as reading the landing page.

3.5.1 Search Effort. We quantify participants' effort in searching, using the following measures:

- **#Queries:** the number of issued queries;
- **#Clicks:** the number of clicked results on SERPs;
- **#Pages:** the number of visited Web pages, including the SERPs and landing pages;
- **Time on SERP:** the time spent on browsing the SERPs;
- **Time on landing page:** the time spent on reading the landing pages;
- **Task time:** the total amount of time spent on completing the search task (i.e. the sum of **Time on SERP** and **Time on landing page**).

Because domain expertise may help the user in completing in-domain search tasks, we hypothesize that domain expert users are more efficient in the in-domain search tasks. For example, we expect that the domain expert user to spend less time on the in-domain tasks. However, it is also possible that the domain expert users will issue more queries and click more results in a in-domain session, because: 1) they may be more interested in the in-domain tasks; and 2) it may take less effort for them to issue queries and judge results when completing an in-domain task.

3.5.2 Query Reformulation.

In exploratory search, the user often issues multiple queries in a search session. Previous research showed that the success of the search session to a large extent depends on the query reformulation behaviors [33] and the domain knowledge level has an effect on the diversity of query vocabulary [26]. Therefore, we are interested in how experts and non-experts reformulate their queries.

We first examine a variety of query reformulation measures, including:

- **#Terms per q.:** the number of terms in a query;
- **#Unique terms:** the number of unique terms in a session;
- **#Unique terms per q.:** $\frac{\texttt{\#Unique terms}}{\texttt{\#queries}}$, the number of unique terms per query (Query vocabulary richness, QVR, in [26]);
- **%Terms from desc.:** the ratio of terms from the task description;
- **%Generalization:** the ratio of generalization reformulation;
- **%Specification:** the ratio of specification reformulation;
- **%Substitution:** the ratio of substitution reformulation.

The Generalization, Specification, and Substitution reformulation were defined by Lau and Horvitz [23]. In this study, the types of reformulations were automatically determined by the following criteria. For a query reformulation from q_0 to q_1, we use S_0 and S_1 to denote the set of terms in q_0 and q_1, respectively. A query reformulation was classified as a Generalization reformulation if $S_1 \subset S_0$, a Specification if $S_1 \supset S_0$, and a Substitution if there exists two terms t_0 and t_1 such that $(t_0 \in S_0 \wedge t_0 \notin S_1)$ and $(t_1 \notin S_0 \wedge t_1 \in S_1)$.

White et al. [37] showed that domain expert users will issue longer queries in the in-domain tasks. Liu et al. [26] further demonstrated that the domain expert users use more diverse terms in queries, especially in difficult tasks. By inspecting the query reformulation measures, we further test whether these effects can be identified across different knowledge domains.

By using an eye-tracker in the user study, we can analyze the origins of *the novel query terms* that are not in the task descriptions. In previous research, Eickhoff [14] used eye-tracking devices to estimate which terms were read by the user and showed that the read terms are likely to be the source for future query reformulation. We also capture the terms that were read by the participants during searching using the eye fixation sequences logged by the eye-tracker. For each visited page, we compute the fixation time $F(t)$ for term t:

$$F(t) = \sum_{e} \sum_{f \in Fixations(e)} \frac{TF(t, Content(e))}{|Content(e)|} \times Duration(f) \tag{2}$$

Here e is an HTML element on page. $Fixations(e)$ is a function that returns all the fixations f that are in the bounding box of e and not in the bounding box of any other HTML element inside e. $Content(e)$ returns the textual content (i.e. a list of terms) of element e. Stop words are filtered from the content because they are rarely processed in reading and we are not interested in the origins of the stop words in queries. $TF(t, Content(e))$ computes the term frequency of t in the text content of element e. $Duration(f)$ is the duration of fixation f.

We then set a threshold for $F(t)$ to extract terms that were actually read by the participant on every page and examine whether the novel query term is from visited SERPs, landing pages, or from other sources (e.g. participants' prior knowledge). Because an HTML element can contain multiple terms (the median of the number of terms in a fixated element is 39 in our dataset) and users often only processes a small proportion of these terms, it may fail to identify which term is actually processed by each fixation. However, the proposed method can effectively filter out the terms in the HTML elements that have no fixations. The computed term-level fixation time $F(t)$ can be regarded as an approximate indicator of the probability that a term is processed by the participant. Due to this limitation, while Reingold et al. [34] suggest a lexical processing threshold of ~145ms, we use a more inclusive threshold of 100ms in our study to get a higher recall of query terms in the fixated HTML elements. Eickhoff et al.'s study [14] also showed that the fixation time on each term varied according to its complexity or familiarity measured by the average age of acquisition (AOA). However, because of a lack of AOA dataset for Chinese words, we used a unified threshold for all terms. We acknowledge this as a limitation of the study.

Based on the fixation time $F(t)$ we compute the following measures for the origins of novel query terms:

- **%From landing page:** the proportion of novel query terms from reading landing pages;
- **%From SERP:** the proportion of novel query terms from reading SERPs;
- **%Others:** the proportion of novel query terms from other sources (i.e. not acquired during search).

We hypothesize that the domain expert users will use more query terms from their prior domain knowledge while the non-expert users will rely more on the query terms acquired through search.

3.5.3 SERP Examination and Clicking. For the examination behaviors on SERPs, we use 100ms of accumulated fixation time as the minimal requirement for result examination and inspect the following dependent variables:

- **#Examined results:** the number of examined results;
- **Exam. time per r.:** the average fixation time on each examined result;
- **Avg. Examined rank:** the average rank of examined results;
- **Max. Examined rank:** the rank of the lowest examined result in the SERP.

For the click actions on SERPs, we also inspect the following measures:

- **#Uniq. clicks per q.:** the average number of unique clicked results per query;
- **P(click|examine):** the average click through rate given the result is examined by the participant;
- **Avg. clicked rank:** the average rank of clicked results;
- **Max clicked rank:** the rank of the lowest clicked result in the SERP;
- **Avg. usefulness:** the average usefulness feedback of clicked results;
- **%useful clicks:** the proportion of clicked results that were marked as useful by the participant;
- **Avg. dwell time:** the average dwell time of clicked results;
- **%SAT clicks:** the proportion of clicked results that have a dwell time longer than 30s [17].

In previous work, Cole et al. [10] found that the user with a high domain knowledge level tends to click top-ranked results because they are more effective in query reformulation and is better at discriminating documents with different knowledge levels. Therefore, we hypothesize that, comparing with non-expert users, the domain expert users: 1) will examine and click shallower results; 2) are better at identifying useful results on SERPs.

3.5.4 Landing Page Reading. The eye-tracking data enable us to analyze the influence of domain expertise on the information acquisition actions on landing pages.

In a previous study, Cole et al. [9] showed that, in the domain of genomics, users' eye movement patterns are associated with their prior knowledge levels. In this study, we test whether this finding can be generalized across domains. We first use the model proposed by Buscher et al. [7] to label the

Table 8. Average relative deviation of the results of the pre-task questionnaire. */**/*** indicate the difference between IN-domain sessions and OUT-domain sessions is significant at $p < 0.05/0.01/0.001$

User domain:	All domains			Environment			Medicine			Politics		
Task sessions:	IN	OUT	Sig.	IN	OUT	Sig.	IN	OUT	p	IN	OUT	Sig.
Pre-knowledge	+32.8%	-15.8%	***	+46.8%	-14.3%	***	+37.3%	-12.7%	***	+14.3%	-20.6%	**
Pre-difficulty	-13.7%	+6.6%	***	-22.1%	-1.0%	*	-3.9%	+16.4%	**	-15.0%	+5.2%	**
Pre-interest	+10.9%	-5.3%	**	+11.6%	-7.6%	*	+16.8%	-5.5%	**	+4.3%	-2.4%	-

sequences of fixations as more engaged *reading* sequences and less engaged *skimming* sequences and compute the following measures:

- **#Fixation per page:** the number of fixations in the landing page;
- **%Reading:** the percentage of reading fixations;
- **%Skimming:** the percentage of skimming fixations;

The percentages of reading and skimming fixations (**%Reading** and **%Skimming**) characterize the attention and effort in reading the landing page.

Then, we compute the average values of five cognitive effort measures proposed by Cole et al. [9] for all the reading sequences:

- **Avg. read length:** the average length of reading sequences measured in the amount of text processed[2];
- **Avg. LADE:** the average LADE (lexical access duration excess, the additional fixation duration beyond minimum time to acquire the meaning of a word);
- **Avg. #regressions:** the average number of regressions in the reading sequence;
- **Avg. perceptual span:** the average horizontal span of the reading sequence;
- **Avg. reading speed:** the ratio of reading length to the processing time.

Note that in this study, instead of using the original reading model in [9], we adopt a simpler yet robust reading model proposed by Buscher et al. [7]. These measures are designed to quantify the cognitive effort of users in reading landing pages. We expect to see domain expert users put less cognitive effort in in-domain tasks.

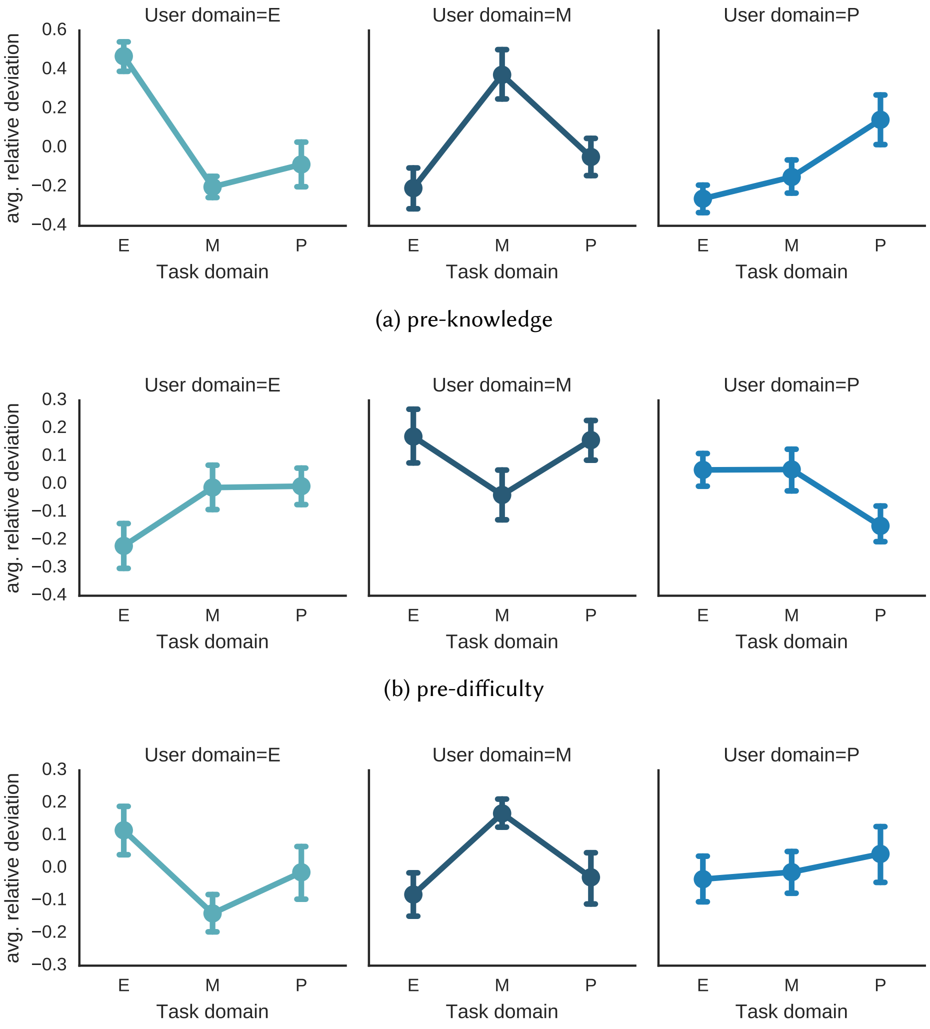
4 EXPERIMENT RESULTS

4.1 Pre-task Questionnaire Results

We first inspect the feedbacks from the pre-task questionnaire to test the effectiveness in controlling participants' domain expertise level. From the average relative deviations of the the pre-task questionnaire results listed in Table 8, we can see that: 1) the participants reported that they have higher domain knowledge levels for in-domain tasks than out-domain tasks; 2) the participants anticipated that the in-domain tasks are easier than the out-domain ones; 3) the participants were generally more interested in the in-domain tasks except for the difference being not significant in the politics domain (User Domain=P). These results confirm our intuition and validate that our experiment design can indeed control the participants' prior domain expertise levels in completing the simulated search tasks.

In Figure 4, we show the average relative deviation of dependent variables (along with the standard errors) in the pre-task questionnaire of the search sessions that belongs to different user domains and task domains. From this figure, we spot some differences across domains. First, although the participants from the politics domain (User Domain=P) reported the highest pre-knowledge value for the politics search tasks (Task Domain=P), the difference of this measure across

[2] We approximately estimate **Avg. read length** by dividing the total left-to-right span of reading sequence by the width of a character



(a) pre-knowledge

(b) pre-difficulty

(c) pre-interest

Fig. 4. Avg. relative deviation of the pre-task questionnaire results in different user and task domains

user domains is relatively small as compared to the pre-knowledge values for the environment and medicine search tasks. A similar pattern can be identified for the pre-interest measures. A possible reason for this is that the politics search tasks are less technical and are more likely to be part of common knowledge, so the participants from different domains reported comparable prior knowledge level and interest level for these tasks. Second, the expected difficulty levels of the medicine search tasks are similar across user domains. There are two possible explanation for this phenomenon: 1) the search tasks in the medicine domain are intrinsically harder than other search tasks, even for the domain expert users in the medicine domain; 2) the participants from the medicine domain might have higher standards in completing the medical search tasks.

Table 9. Average relative deviation of the search outcome measures. */**/*** indicate the difference between IN-domain sessions and OUT-domain sessions is significant at $p < 0.05/0.01/0.001$

User domain:	All domains			Environment			Medicine			Politics		
Task sessions:	IN	OUT	Sig.	IN	OUT	Sig.	IN	OUT	Sig.	IN	OUT	Sig.
Post-knowledge	+9.7%	-4.7%	***	+6.4%	-6.2%	**	+10.0%	-4.4%	*	+12.7%	-3.3%	**
Answer score	+8.0%	-3.9%	*	+17.8%	+1.6%	-	+9.9%	-4.3%	*	-3.6%	-9.4%	-
Post-difficulty	-6.0%	+2.9%	*	-15.8%	-10.1%	-	+9.8%	+9.7%	-	-11.9%	+10.5%	*
Post-interest	+7.4%	-3.5%	*	+5.8%	-4.0%	-	+12.0%	-3.5%	-	+4.3%	-3.1%	-
Satisfaction	+5.3%	-2.6%	-	+2.9%	-0.8%	-	+0.7%	-7.0%	-	+12.3%	-0.1%	*

Table 10. Average relative deviation of the search effort measures. */**/*** indicate the difference between IN-domain sessions and OUT-domain sessions is significant at $p < 0.05/0.01/0.001$

User domain:	All domains			Environment			Medicine			Politics		
Task sessions:	IN	OUT	Sig.	IN	OUT	Sig.	IN	OUT	Sig.	IN	OUT	Sig.
#Queries	-5.3%	+2.5%	-	+3.8%	+1.3%	-	-6.2%	-10.5%	-	-13.4%	+17.0%	-
#Clicks	-6.4%	+3.1%	-	-10.1%	-4.7%	-	+1.7%	+12.8%	-	-10.7%	+1.9%	-
#Pages	-7.1%	+3.4%	-	-1.1%	+1.7%	-	-1.0%	+7.0%	-	-19.2%	+1.8%	-
Time on SERP	-11.3%	+5.4%	*	-3.5%	-0.5%	-	-13.8%	-1.5%	-	-16.5%	+18.9%	*
Time on landing page	-7.1%	+3.4%	-	-14.1%	-9.1%	-	-11.3%	+2.3%	-	+4.2%	+18.4%	-
Task time	-7.8%	+3.8%	*	-12.6%	-7.8%	-	-10.4%	+4.2%	-	-0.5%	+16.3%	-

4.2 Search Outcomes

We first examine the effect of domain expertise on two aspects of search outcome, search success and search satisfaction. As we mention in Section 3.4, the search success is measured by the subjective post-knowledge feedback and objective answer score, while the search satisfaction is measured the participants' feedback on satisfaction, perceived difficulty, and increase of interest.

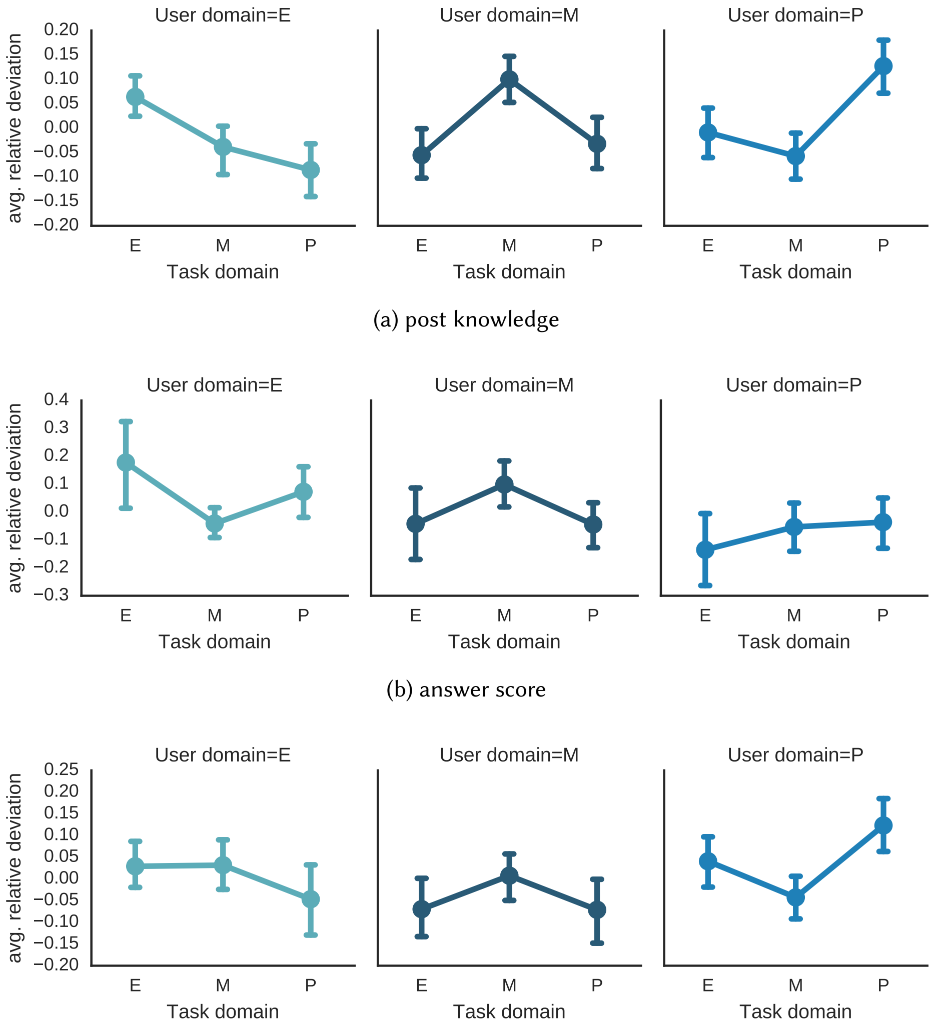
From Table 9, we can see that: 1) For search success, the participants from all three domains reported higher subjective post-knowledge levels for in-domain tasks; although only the difference in the medicine domain is statistically significant, the objective answer scores of in-domain tasks were consistently higher than those of out-domain tasks. These findings indicate that the participants were more successful in the in-domain search tasks. 2) For search satisfaction, although the participants felt that the in-domain tasks were more interesting (post-interest) and easier to complete (post-difficulty), only the participants from the politics domain were significantly more satisfied with the search process of the in-domain tasks.

We further inspect the differences in search outcomes for three user domains. Reflected by the relative deviations shown in Figure 5, the participants from the environment domain have a relatively high average answer score, while the participants from politics domain have a low average answer score. This difference might be caused by the difference in the search expertise of the participants with varied backgrounds.

These results suggest that the domain expertise has a consistent effect on the success of exploratory search, while its effect on search satisfaction is conditioned on the user domains.

4.3 Search Process

4.3.1 Search Effort. From Table 10, we can see that although only the differences of task time and time on SERP are significant, all the search effort measures suggest the participants put less effort in completing in-domain search tasks. These results contrast with White et al. [37]'s results that the domain experts may issue more queries, visit more pages, and spend more time in the in-domain search sessions. In our lab-based study we control the search tasks, therefore, the domain experts completed the same search tasks more efficiently than the non-experts. The comparisons of user behavior measures are not between the sessions of the same search task, as in White et



(a) post knowledge

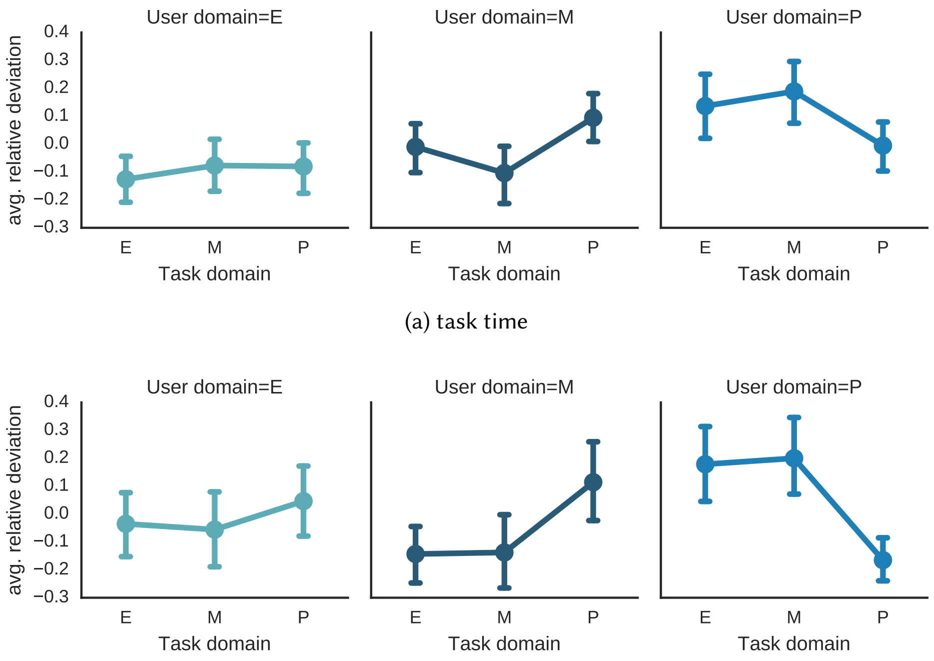
(b) answer score

(c) satisfaction

Fig. 5. Avg. relative deviation of the search outcome measures in different user and task domains

al. [37]. In their naturalistic log-based study, the in-domain search tasks completed by the domain experts may be intrinsically more complex and difficult, and therefore require more search effort.

It is also interesting to see in Figure 6 that the participants from the politics domain (User Domain=P) spent much less task time and time on SERP in the in-domain tasks (Task Domain=P) than in the out-domain tasks (Task Domain=E and M). Combined with the fact that the politics users had a relatively low answer score and high satisfaction for the in-domain tasks (see Section 4.2), it seems that a high domain expertise level made the participants from politics domain not more effective but more efficient in the in-domain search tasks. A high level of efficiency (especially in examining the SERPs) might explain why they felt more satisfied with the search processes of in-domain sessions.



(a) task time

(b) time on SERP

Fig. 6. Avg. relative deviation of the search effort measures in different user and task domains

Table 11. Average relative deviation of query reformulation measures. */**/*** indicate the difference between IN-domain sessions and OUT-domain sessions is significant at $p < 0.05/0.01/0.001$

User domain:	All domains			Environment			Medicine			Politics		
Task sessions:	IN	OUT	Sig.	IN	OUT	Sig.	IN	OUT	Sig.	IN	OUT	Sig.
#Terms per q.	+3.7%	-1.8%	-	-2.6%	-7.1%	-	-7.0%	-4.8%	0.223	+20.6%	+7.2%	-
#Unique terms	+0.0%	-0.0%	-	-5.4%	-10.2%	-	-12.0%	-11.0%	0.381	+17.4%	+22.4%	-
#Unique terms per q.	+6.6%	-3.2%	-	-13.4%	-10.0%	-	-5.4%	-8.9%	0.395	+38.7%	+10.1%	-
%Terms from desc.	-0.3%	+0.1%	-	-2.6%	+1.4%	-	+1.8%	+0.1%	0.360	-0.0%	-1.3%	-
%Generalization	-0.7%	+0.3%	-	+46.2%	+8.4%	-	-44.1%	-14.5%	0.236	-4.2%	+6.2%	-
%Specification	-1.6%	+0.8%	-	-19.5%	-2.0%	-	+23.3%	+16.7%	0.443	-8.6%	-12.1%	-
%Substitution	+27.1%	-13.1%	-	-6.9%	-20.5%	-	+78.6%	-11.1%	0.222	+9.7%	-6.7%	-

Table 12. Average relative deviation of the sources of novel query terms. */**/*** indicate the difference between IN-domain sessions and OUT-domain sessions is significant at $p < 0.05/0.01/0.001$

User domain:	All domains			Environment			Medicine			Politics		
Task sessions:	IN	OUT	Sig.	IN	OUT	Sig.	IN	OUT	Sig.	IN	OUT	Sig.
%From landing page	-17.9%	+8.6%	**	-32.6%	-0.1%	*	-6.6%	+15.7%	-	-14.5%	+11.2%	-
%from SERP	-13.4%	+6.5%	*	+0.0%	+7.7%	-	-37.4%	-11.5%	*	-2.9%	+23.1%	-
%Others	+8.9%	-4.3%	*	+3.2%	-5.2%	-	+17.6%	+4.7%	-	+5.9%	-12.2%	-

4.3.2 Query Reformulation. We first examine a variety of query reformulation measures, including the the number of terms in the query (**#Terms per q.**), the number of unique terms in the session (**#Unique terms** and **#Unique terms per q.**), the proportion of the terms that are in task descriptions (**%terms from desc.**), and the proportion of different query refinements used

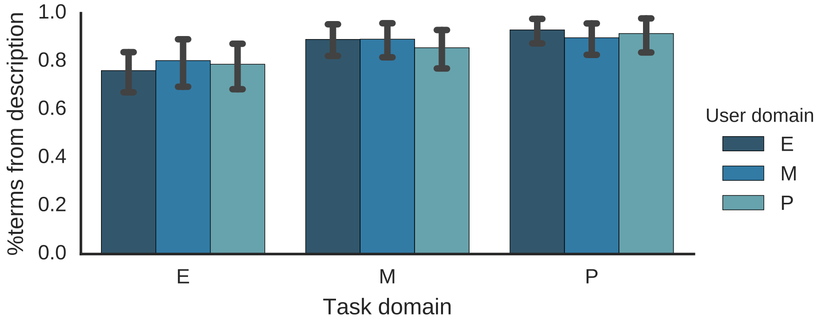


Fig. 7. %query terms in task description

Table 13. Average relative deviation of the SERP examination behavior measures. */**/*** indicate the difference between IN-domain sessions and OUT-domain sessions is significant at $p < 0.05/0.01/0.001$

User domain:	All domains			Environment			Medicine			Politics		
Task sessions:	IN	OUT	Sig.	IN	OUT	Sig.	IN	OUT	Sig.	IN	OUT	Sig.
#Examined results	-8.4%	+4.0%	-	-8.2%	+4.0%	-	-1.2%	+16.2%	-	-15.7%	-8.0%	-
Exam. time per r.	+11.3%	-5.5%	-	-14.4%	-23.3%	-	+33.5%	+1.1%	-	+14.9%	+7.7%	-
Avg. examined rank	-10.8%	+5.2%	-	-5.0%	+12.0%	-	-2.8%	+15.5%	-	-24.6%	-12.6%	-
Max examined rank	-8.5%	+4.1%	-	-4.7%	+4.4%	-	+1.6%	+21.4%	-	-22.3%	-13.6%	-

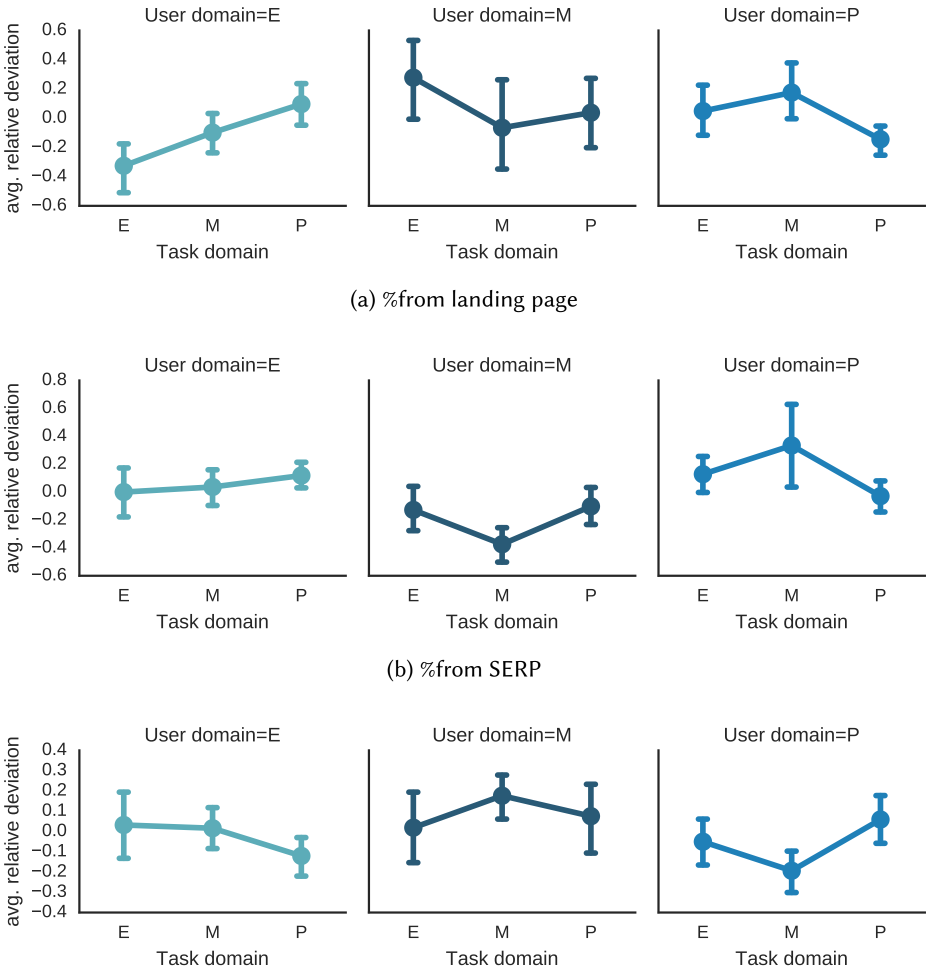
(**%Generalization**, **%Specification**, and **%Substitution**). We list the results in Table 11. To our surprise, we did not find any significant effect of domain expertise on these measures. Upon further inspection of the participants' query terms we find that 86.1% of the query terms come from the task description (Figure 7). This may explain why the query reformulation behaviors in the in-domain and out-domain search sessions are similar in our dataset.

Therefore, we focus on the origins of the novel query terms that are not included in the task descriptions. From the results presented in Table 12, we can see that the participants from all user domains used more query terms that were from landing pages or SERPs during the out-domain sessions; and they used more terms that were not read on visited pages in the in-domain sessions. This finding suggests the domain expertise affects how the user selects query terms during the exploratory search. When exploring a new domain, the user may accumulate vocabulary and learn how to query during the search. When performing in-domain search tasks, the user may have enough prior knowledge to come up with effective query terms.

From Figure 8, we further find that when completing the medicine search tasks, the participants from the medicine domain use more terms from other sources than other participants. This finding suggests that the background knowledge is more important in formulating good search queries in a highly technical knowledge domain like the medicine domain.

4.3.3 SERP Examination and Clicking. After examining the query reformulation behaviors, we focus on how the participants with different domain expertise levels examine and click the results on SERPs.

For the examination behaviors, from Table 13 we can see that, while none of the difference is statistically significant, the participant examined fewer (**#Examined results**) and shallower results (**Avg. examined rank** and **Max examined rank**) but spent more time in examining a single search result (**Exam. time per r.**) in the in-domain sessions. Some click-related measures (**Avg. clicked rank** and **Max clicked rank**) in Table 14 reveal a similar tendency for the domain



(a) %from landing page

(b) %from SERP

(c) %others (e.g. prior knowledge)

Fig. 8. Avg. relative deviation of the sources of novel query terms in different user and task domains

Table 14. Average relative deviation of the SERP click behavior measures. */**/*** indicate the difference between IN-domain sessions and OUT-domain sessions is significant at $p < 0.05/0.01/0.001$

User domain:	All domains			Environment			Medicine			Politics		
Task sessions:	IN	OUT	Sig.	IN	OUT	Sig.	IN	OUT	Sig.	IN	OUT	Sig.
#Uniq. clicks per q.	-1.5%	+0.7%	-	-17.8%	-7.5%	-	+20.2%	+9.7%	-	-6.8%	+0.9%	-
P(click|examine)	+6.9%	-3.3%	-	-12.3%	-14.8%	-	+21.4%	+0.9%	-	+11.6%	+5.2%	-
Avg. clicked rank	-5.8%	+2.8%	-	-7.1%	+0.7%	-	+4.9%	+12.7%	-	-15.3%	-4.8%	-
Max clicked rank	-10.5%	+5.1%	-	-14.7%	+1.5%	-	+0.3%	+20.5%	-	-17.0%	-6.5%	-
Avg. usefulness	-4.1%	+2.0%	-	-8.3%	+7.4%	*	-3.5%	+1.2%	-	-0.4%	-3.3%	-
%Useful clicks	-11.8%	+5.7%	-	-17.2%	+24.8%	-	-10.7%	+4.7%	-	-7.6%	-14.4%	-
Avg. dwell time	-1.0%	+0.5%	-	-11.3%	-7.9%	-	-12.5%	-6.9%	-	+20.8%	+17.2%	-
%SAT clicks	-11.0%	+5.3%	*	-21.4%	+1.0%	-	-22.5%	-10.7%	-	+11.0%	+26.1%	-

Table 15. Average relative deviation of the landing page reading behavior measures. */**/*** indicate the difference between IN-domain sessions and OUT-domain sessions is significant at $p < 0.05/0.01/0.001$

User domain:	All domains			Environment			Medicine			Politics		
Task sessions:	IN	OUT	Sig.	IN	OUT	Sig.	IN	OUT	Sig.	IN	OUT	Sig.
#Fixations per page	+0.6%	-0.3%	-	-12.1%	-8.5%	-	-2.0%	-2.3%	-	+16.0%	+10.7%	-
%Reading	-3.8%	+1.9%	-	-7.7%	-1.2%	-	+0.1%	+6.9%	-	-3.9%	+0.3%	-
%Skimming	+8.0%	-3.9%	-	+18.2%	+7.4%	-	-21.4%	-32.6%	-	+27.3%	+12.4%	-
Avg. read length	-8.8%	+4.2%	-	-16.8%	-2.6%	-	+4.5%	+23.4%	-	-14.0%	-7.3%	-
Avg. LADE	-10.2%	+4.9%	*	-29.8%	-9.1%	-	-6.6%	+14.2%	-	+5.9%	+11.2%	-
Avg. #regressions	+0.7%	-0.4%	-	-20.7%	-25.6%	-	+37.7%	+32.6%	-	-14.8%	-5.3%	-
Avg. perceptual span	-3.7%	+1.8%	-	+0.3%	+8.4%	-	-9.3%	-4.8%	-	-2.2%	+1.0%	-
Avg. reading speed	+4.5%	-2.2%	*	+5.9%	+4.5%	-	+4.9%	-1.9%	-	+2.7%	-9.9%	*

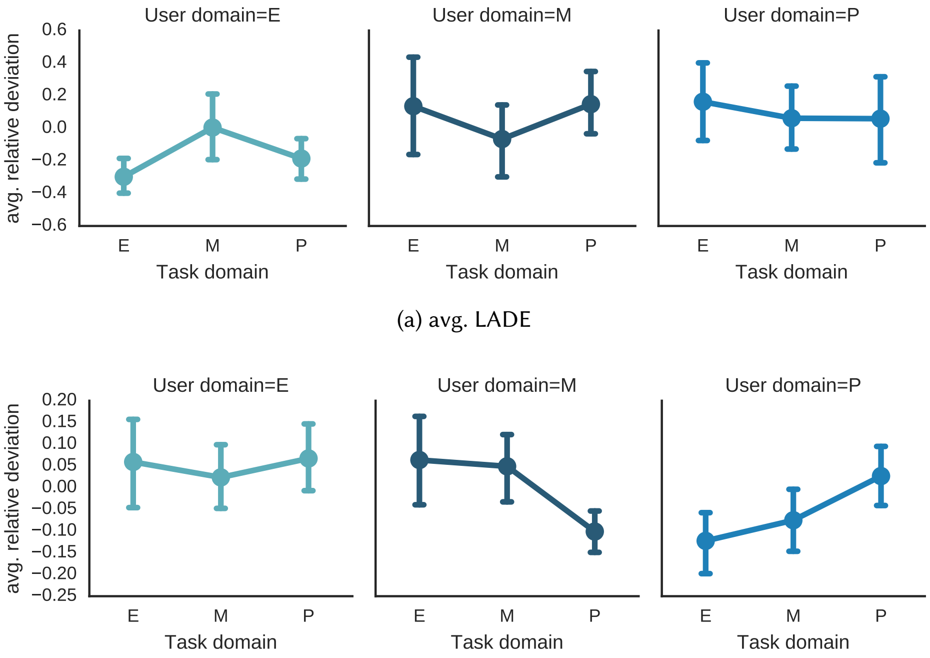
expert users to click top-ranked results. These findings are consistent with Cole et al.'s findings in previous work [10] that the users with a high domain knowledge level exhibit a slightly stronger position bias in their clicks on SERPs. A potential reason for this is that the domain expert users issue better queries than the non-expert users, therefore, they are more likely to find relevant results at higher rank.

We also hypothesize that, comparing with non-expert users, the domain expert users are better at identifying useful results on SERPs, thus their clicks are associated with higher usefulness feedbacks and longer dwell time. However, the results support the opposite of our intuitive hypothesis. The clicked results in the in-domain sessions seems to be less useful than the clicked results in the out-domain sessions. The proportion of SAT clicks (**%SAT clicks**) is lower in the in-domain sessions than in the out-domain sessions. The **Avg. usefulness** value for the in-domain sessions conducted by the participants from environment domain (User Domain=E) is significantly lower than those for the out-domain sessions. These findings suggest that, instead of clicking better results on SERPs, the domain expert users may have a higher usefulness criterion when accessing the landing pages. Therefore, they are more likely to leave a landing page within the 30s time threshold and their explicit usefulness feedbacks for the landing pages are lower.

4.3.4 Landing Page Reading. From the results in Table 15, we find that the participants put less cognitive effort in the in-domain search sessions. The **Avg. LADE** values are consistently lower in the in-domain sessions across all three user domains. If we combine all the user domains together, the difference is statistically significant. The **Avg. reading speed** values are consistently higher in the in-domain search sessions and the difference is also significant in the politics domain (user domain=P) and on the whole dataset. From Figure 9 we find that the cognitive effort measures vary across different user domains. The participants from the environment domain (User domain=E) have relatively high reading speed and low LADE value while the participants from the politics domain have low reading speed and high LADE value. The participants from the medicine domain have a very low reading speed when completing the unfamiliar politics tasks. We generalize Cole et al.'s [9] findings to new knowledge domains and find that while the domain expertise level has domain-independent effects on the cognitive effort measures derived from eye movement patterns, the cognitive effort measures themselves depend on the knowledge domains.

5 DISCUSSION

Understanding how domain expertise level affects the process and outcome of exploratory search is crucial for improving the search engine in supporting such complex search tasks. On the one hand, understanding the effect on search outcome can reveal and explain why some searches are successful or satisfying for users while other searches are not. This is important for the evaluation and failure analysis of search systems. On the other hand, the analysis of the effect on user's search



(a) avg. LADE

(b) avg. reading speed

Fig. 9. Avg. relative deviation of the measures of landing page reading behaviors in different user and task domains

behavior can provide guidance in developing methods and models that can estimate user's domain knowledge level and enhance the personalization of search system.

In order to fulfill this objective, we conducted a dedicated lab-based user study to investigate the effect of domain expertise level in exploratory search. In the user study we controlled the independent variable, user's domain knowledge level, by designing simulated search tasks in three different knowledge domains and hiring participants with different backgrounds to complete them. We regard the participants working with in-domain tasks as domain expert users with a higher level of domain knowledge and the participants working with out-domain tasks as non-expert users. We acknowledge that this binary domain knowledge level is different from the continuous domain knowledge level measured by domain-specific thesaurus [40] and quizzes [12], or operationalized by longitudinal user studies [36, 39]. However, by assessing the self-reported pre-knowledge, expected difficulty and interest level in the pre-task questionnaire, we validated the experimental manipulation of the domain knowledge level of the participants in the study.

The advantages of this study's experimental design include: 1) We can fully control the search tasks and directly compare the search sessions conducted by domain expert users and non-expert users to investigate the effect of domain expertise level. 2) It is easier to generalize the approach to different knowledge domains because no domain-specific thesaurus or quizzes are needed.

There are also some limitations with the experiment settings, such as: 1) As shown by the gray line in Figure 1, the search tasks and user domains can also affect the dependent variables (i.e. measures for search outcome and search process). Therefore, we have to use the method described in Section 3.2 to avoid these confounding effects. 2) There may be cross-domain expertise between different knowledge domains. For example, it is possible that the participants from environment domain are more familiar with the tasks from medicine domain than those from politics domain. The cross-domain expertise effect may reduce the probability of discriminating in-domain and out-domain sessions. Therefore, it is necessary to use the pre-task questionnaire to validate the experiment control and prevent adopting knowledge domains that are closely related.

Regarding **RQ1**, we inspect how domain expertise affect the search success and search satisfaction. For search success, we find that, measured by the subjective self-reported knowledge gain and objective answer score, the participants from all the three domains are in general more successful in the in-domain tasks than in the out-domain tasks. This result confirms our hypothesis and the findings in previous studies [12, 37], suggesting prior domain knowledge can help users search for in-domain information more effectively. However, although the participants in general rate the in-domain tasks as more interesting and less difficult than the out-domain tasks, only the participants from politics domain are significantly more satisfied in the in-domain tasks. Search satisfaction depends on not only whether the useful and correct information is found during the search, but also other factors like the search effort and user's expectation of search results. Some recent studies [24, 27] characterize the difference between search satisfaction and search success. Our findings also indicate that the effects of domain expertise level on search success and search satisfaction are different and it is interesting to further analyze why the domain expert users are not satisfied with some successful exploratory search sessions in future work.

Regarding **RQ2**, we investigate the effect of domain expertise on a variety of search behavior measures. We first find that the participants are more efficient in completing the in-domain search tasks, reflected by a significant difference in **Task time** measures and a consistent decrease of other search effort measures. These results seem contrary to White et al.'s findings that domain experts issue more queries, visit more pages, and spend more time on the in-domain tasks [37]. This disagreement in results might be due to the difference in the experiment settings of these two studies, in particular, whether the search tasks are held constant across participants. These differences in findings also emphasize the advantage of the adopted experiment settings and how this study complements the existing research literature on the effect of domain expertise in search.

We expected to see the difference in query reformulation patterns between domain expert users and non-expert users. However, since the scale of the study is relatively small and most of the query terms used by the participants are from the corresponding task descriptions, we failed to identify any statistically significant difference in user's query reformulation patterns characterized by the measures in Table 11. With an eye-tracker and the Chrome extension that logs all the content of visited pages, we captured the text content actually read by the participants during the experiment and use the fixation time $F(t)$ of each novel query term as a probabilistic indicator of its source. By analyzing the source of novel query terms, we confirm that the domain expert users use more terms that are not acquired during search and the non-expert users tend to build their query vocabulary during search. This new finding provides direct evidence for the hypothesis that with the help of prior domain knowledge, the domain expert users are better at issuing queries in exploratory search. It also suggests that the success in exploratory search may depends on whether the user can issue more effective query terms, which emphasizes the importance of providing good query suggestions for highly technical domains such as the medicine domain.

The results on SERP examination and clicking behavior is consistent with Cole et al.'s findings [10] that the clicks from the user with high domain knowledge level are slightly more biased toward top-ranked results. It is also interesting to see that the results clicked by the domain expert users are not necessarily more useful. A possible explanation for this phenomenon is that the domain expert user may have a higher standard for result usefulness and can make usefulness judgment in shorter time, which results in a lower usefulness feedback (**Avg. usefulness** and **%Useful clicks**) and a lower percentage of SAT clicks (**%SAT clicks**). These findings may provide implications for how to better utilize the query log of domain expert users [37]. Because the results clicked by the domain expert users are not necessarily more useful, we should not discriminate the importance of expert users' click and normal users' click. However, because the domain expert users have a higher standard for usefulness, their SAT clicks may be strong indicators for high quality results.

The results of reading behavior on landing pages show that it takes less cognitive effort for the domain expert users to read landing pages in the in-domain tasks, generalizing Cole et al.'s [9] findings to domains other than medicine and biology. The **Avg. LADE** measure is associated with the proportion of technical terms that takes more cognitive effort to process. One would expect that when the non-expert user is completing the tasks in medicine domain, the **Avg. LADE** will be larger by a significant margin. However, the results in Figure 9 do not correspond to this hypothesis. A possible reason for this is that unlike in English many medical terms in Chinese are composed by common characters, which will not be hard to process for the participants.

Regarding **RQ3**, we also find some domain-specific effect of user's domain expertise level. For example, we find that the participants from the politics domain put less effort and feel more satisfied in the in-domain tasks. The participants from medicine domain report high expected difficulty for both in-domain and out-domain tasks and use more novel query terms from other sources. The existence of domain-specific effect implies that some findings in previous studies that only focus on a specific domain may not generalize to other knowledge domains and it is necessary to compare effect across multiple domains to validate whether the effect is domain-independent or not. The difference of domain expertise effect across different domains also suggests that, in order to personalize the search results according to user's domain knowledge level, we need to develop different models in different knowledge domains.

Finally, we acknowledge some limitations of our study. First, as in most user studies, the number of participants is limited. Our dataset may not have enough statistical power to identify some subtle effects of domain expertise level, especially the effects conditioned on the user or task domains. Second, all the participants in this study were college students, so they may not be representative of the real Web search engine user group. To overcome these weaknesses, a user study or a crowdsourcing-based study with larger scales and wider coverage of different user background is needed in the future. Third, we use predefined search tasks in the user study. Although we carefully designed the search tasks to simulate practical learning-related search scenarios, the settings may alter the search behaviors. For example, while a real user may abandon the search if she can not find some relevant information at the beginning of the search, the participant in the user study may choose to continue searching in the same situation.

6 CONCLUSIONS

In this work, we study the effects of domain expertise levels on the process and outcome of exploratory search. Compared to existing research, we 1) propose a new experiment setting to control participants' domain knowledge level in a lab-based user study; 2) use over 40 measures to characterize the effect of domain knowledge level on the process and outcome of exploratory search, which complement and extend the findings in existing research; 3) investigate and compare the effects of domain knowledge level across three different domains, allowing us to identify the domain-independent and domain-specific effects.

Our analysis confirms that a high domain expertise level often leads to a higher success rate in completing the search tasks but we fail to detect a domain-independent domain expertise effect on user's satisfaction (**RQ1**). We investigate a series of user behavior measures (**RQ2**) and have the following observations: 1) The domain expert users can complete in-domain search tasks more efficiently; 2) With task descriptions being the major source for query vocabulary, the participants may use more new query terms from landing pages and SERPs when exploring unfamiliar domains; 3) Measured by dwell time and explicit usefulness feedback, the results clicked by the domain experts are not necessarily more useful; 4) The domain experts put less cognitive effort in reading the landing page. Besides identifying some domain-independent effects, we also find some effects in particular domains (**RQ3**), including: 1) The participants from the politics domain put less effort and feel more satisfied in completing the in-domain tasks; 2) The participants from the medicine domain use more novel query terms that are probably from their prior knowledge in their domain of expertise.

These findings may provide useful implications for the design of search systems. For instance: 1) We find that the results clicked by domain experts may not be more relevant (Section 4.3.3). Therefore, we should re-think about how to exploit domain experts' click logs. 2) With an eye-tracker, we find that the non-expert users use more terms encountered during search as their query terms, especially in the medicine domain (Section 4.3.2). We should enhance the query suggestion function for highly technical domains like the medicine domain because exploring such domains may require domain-specific query vocabulary. 3) Regarding **RQ3**, because the effects of domain expertise level on user's search behavior may be different in different knowledge domain, when trying to personalize the search results according to user's domain knowledge level, we need to develop different models for different knowledge domains.

APPENDIX

Table 16. The search tasks adopted in the user study (Table 1 in Chinese).

Domain	Task ID	Task Description
Environment	E_1	问：请问我国颗粒物污染（简称PM）特征有哪些？请从全国、地区层面，时间变化层面、颗粒物组成层面等角度进行分析。
	E_2	问：饮用水消毒工艺中紫外消毒不能完全取代氯消毒的原因？
Medicine	M_1	问：目前临床上治疗肿瘤的主要方法及其各自的优缺点？
	M_2	问：3D打印对于精准医疗有哪些可能的应用？
Politics	P_1	问：政治学者注意到，美国大选中党派极化的趋势日益明显，其背后的原因有什么？（极化是指政治观点从中间向两端分散，形成两个敌对的阵营。政党认同更为强烈，更为有力地拒斥另一政党。）
	P_2	问：美国的利益集团为了实现自己的利益，通常会采取那些策略？

Table 17. The questions in the pre-task questionnaire (Table 4 in Chinese)

Measure	Question
Pre-knowledge	你对该搜索任务的主题有多了解？
Pre-difficulty	你预计完成该任务对你来说是否困难？
Pre-interest	你是否对该搜索任务及相关主题感到有兴趣？

Table 18. The questions in the post-task questionnaire (Table 5 in Chinese)

Measure	Question
Post-knowledge	经过搜索，你对该搜索任务的了解增进了多少？
Post-difficulty	经过搜索，你认为该搜索任务是否困难？
Post-interest	经过搜索，你对该搜索任务的兴趣增进了多少？
Satisfaction	对于整个搜索过程，你是否感到满意？

REFERENCES

[1] John R Anderson. 2013. *The architecture of cognition.* Psychology Press.

[2] Kumaripaba Athukorala, Dorota Głowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. 2016. Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2635–2651.

[3] Kumaripaba Athukorala, Antti Oulasvirta, Dorota Głowacka, Jilles Vreeken, and Giulio Jacucci. 2014. Narrow or broad?: Estimating subjective specificity in exploratory search. In *CIKM'14*. ACM, 819–828.

[4] Suresh K Bhavnani. 2001. Important cognitive components of domain-specific search knowledge. *Ann Arbor* 1001 (2001), 48109–1092.

[5] Suresh K Bhavnani. 2002. Domain-specific search strategies for the effective retrieval of healthcare and shopping information. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*. ACM, 610–611.

[6] Pia Borlund. 2000. Experimental components for the evaluation of interactive information retrieval systems. *Journal of documentation* 56, 1 (2000), 71–90.

[7] Georg Buscher, Andreas Dengel, and Ludger van Elst. 2008. Eye movements as implicit relevance feedback. In *CHI'08 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2991–2996.

[8] Jacob Cohen. 1988. Statistical power analysis for the behavioral sciences . Hilsdale. *NJ: Lawrence Earlbaum Associates* 2 (1988).

[9] Michael J Cole, Jacek Gwizdka, Chang Liu, Nicholas J Belkin, and Xiangmin Zhang. 2013. Inferring user knowledge level from eye movement patterns. *Information Processing & Management* 49, 5 (2013), 1075–1091.

[10] Michael J Cole, Xiangmin Zhang, Chang Liu, Nicholas J Belkin, and Jacek Gwizdka. 2011. Knowledge effects on document selection in search results pages. In *SIGIR'11*. ACM, 1219–1220.

[11] Kevyn Collins-Thompson, Soo Young Rieh, Carl C Haynes, and Rohail Syed. 2016. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *CHIIR'16*. ACM, 163–172.

[12] Geoffrey B Duggan and Stephen J Payne. 2008. Knowledge in the head and on the web: Using topic expertise to aid search. In *SIGCHI'08*. ACM, 39–48.

[13] Yuka Egusa, Hitomi Saito, Masao Takaku, Hitoshi Terai, Makiko Miwa, and Noriko Kando. 2010. Using a concept map to evaluate exploratory search. In *IIiX'10*. ACM, 175–184.

[14] Carsten Eickhoff, Sebastian Dungs, and Vu Tran. 2015. An eye-tracking study of query reformulation. In *SIGIR'15*. ACM, 13–22.

[15] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the journey: a query log analysis of within-session learning. In *WSDM'14*. ACM, 223–232.

[16] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[17] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)* 23, 2 (2005), 147–168.

[18] Ahmed Hassan Awadallah, Ryen W. White, Patrick Pantel, Susan T. Dumais, and Yi-Min Wang. 2014. Supporting Complex Search Tasks. In *CIKM'14*. ACM, 829–838.

[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 *SIGIR'14*. ACM, 607–616.

[20] Ruogu Kang and Wai-Tat Fu. 2010. Exploratory information search by domain experts and novices. In *IUI'10*. ACM, 329–332.

[21] Evangelos Kanoulas, Ben Carterette, Mark Hall, Paul Clough, and Mark Sanderson. 2011. Overview of the trec 2011 session track. (2011).

[22] Weize Kong and James Allan. 2014. Extending faceted search to the general web. In *CIKM'14*. ACM, 839–848.

[23] Tessa Lau and Eric Horvitz. 1999. Patterns of search: analyzing and modeling web query refinement. In *UM99 User Modeling*. Springer, 119–128.

[24] Xin Li, Yiqun Liu, Rongjie Cai, and Shaoping Ma. 2017. Investigation of User Search Behavior While Facing Heterogeneous Search Services. In *WSDM'17*. ACM, 161–170.

[25] Yuelin Li and Nicholas J Belkin. 2008. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management* 44, 6 (2008), 1822–1837.

[26] Chang Liu, Xiangmin Zhang, and Wei Huang. 2016. The Exploration of Objective Task Difficulty and Domain Knowledge Effects on Users' Query Formulation. In *ASIST'16*. American Society for Information Science, Article 63, 9 pages.

[27] Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. "Satisfaction with Failure" or "Unsatisfied Success": Investigating the Relationship between Search Success and User Satisfaction. In *WWW'18*. ACM.

[28] Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. 2015. Different users, different opinions: Predicting search satisfaction with mouse movement information. In *SIGIR'15*. ACM, 493–502.

[29] Zeyang Liu, Yiqun Liu, Ke Zhou, Min Zhang, and Shaoping Ma. 2015. Influence of vertical result in web search examination. In *SIGIR'15*. ACM, 193–202.

[30] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.

[31] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. 2016. When Does Relevance Mean Usefulness and User Satisfaction in Web Search?. In *SIGIR'16*. ACM, 463–472.

[32] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[33] Daan Odijk, Ryen W White, Ahmed Hassan Awadallah, and Susan T Dumais. 2015. Struggling and success in web search. In *CIKM'15*. ACM, 1551–1560.

[34] Eyal M Reingold, Erik D Reichle, Mackenzie G Glaholt, and Heather Sheridan. 2012. Direct lexical control of eye movements in reading: Evidence from a survival analysis of fixation durations. *Cognitive psychology* 65, 2 (2012), 177–206.

[35] Lynda Tamine and Cecile Chouquet. 2017. On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings. *Information Processing & Management* 53, 2 (2017), 332–350.

[36] Pertti Vakkari, Mikko Pennanen, and Sami Serola. 2003. Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing & Management* 39, 3 (2003), 445–463.

[37] Ryen W White, Susan T Dumais, and Jaime Teevan. 2009. Characterizing the influence of domain expertise on web search behavior. In *WSDM'09*. ACM, 132–141.

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

[39] Barbara M Wildemuth. 2004. The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology* 55, 3 (2004), 246–258.

[40] 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.

[41] Xiangmin Zhang, Jingjing Liu, Michael Cole, and Nicholas Belkin. 2015. Predicting users' domain knowledge in information retrieval using multiple regression analysis of search behaviors. *Journal of the Association for Information Science and Technology* 66, 5 (2015), 980–1000.

Received September 2017; revised December 2017; accepted May 2018