Investigating User Behavior in Legal Case Retrieval

Yunqiu Shao
BNRist, DCST, Tsinghua University
Beijing, China
shaoyq18@mails.tsinghua.edu.cn

Yueyue Wu
BNRist, DCST, Tsinghua University
Beijing, China
wuyueyue1600@gmail.com

Yiqun Liu*
BNRist, DCST, Tsinghua University
Beijing, China
yiqunliu@tsinghua.edu.cn

Jiaxin Mao
GSAI, Renmin University of China
Beijing, China
maojiaxin@gmail.com

Min Zhang
BNRist, DCST, Tsinghua University
Beijing, China
z-m@tsinghua.edu.cn

Shaoping Ma
BNRist, DCST, Tsinghua University
Beijing, China
msp@tsinghua.edu.cn

ABSTRACT

Legal case retrieval is a specialized IR task aiming to retrieve supporting cases given a query case. While recent research efforts are committed to improving the automatic retrieval models' performances, little attention has been paid to the practical search interactions between users and systems in this task. Therefore, we focus on investigating user behavior in the scenario of legal case retrieval. Specifically, we conducted a laboratory user study that involved 45 participants majoring in law to collect users' rich interactions and relevance assessments. With the collected data, we first analyzed the characteristics of the search process in legal case retrieval practice. We observed significant differences between legal case retrieval and general web search in various search behavior. These differences highlight the necessity of in-depth investigating user behavior in legal case retrieval and re-thinking the application of related mechanisms developed based on the user models in Web search. Then we investigated factors that would influence search behavior from different perspectives, including task difficulty and domain expertise. Finally, we shed light on implicit feedback in legal case retrieval and designed a predictive model for relevance based on user behavior. Our work provides a better understanding of user interactions in the legal case retrieval process, which can benefit the design of the corresponding retrieval systems to support legal practitioners.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; *Specialized information retrieval*.

KEYWORDS

Legal case retrieval, user behavior, implicit relevance feedback

*Corresponding author

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462876

ACM Reference Format:
Yunqiu Shao, Yueyue Wu, Yiqun Liu, Jiaxin Mao, Min Zhang, and Shaoping Ma. 2021. Investigating User Behavior in Legal Case Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3404835.3462876

1 INTRODUCTION

Legal case retrieval is a specialized IR task that involves retrieving supporting cases given a query case. Along with statutes, prior cases decided in courts of law (also called "precedents") are primary legal materials in various law systems. For instance, precedents are fundamental to preparing arguments in common law. In some countries following the civil law system, drawing analogies across relevant prior cases is also increasingly required in pursuit of justice [18]. There have been numbers of benchmark works in recent years, such as TREC legal track [35], COLIEE [36], etc. Recent efforts have been made to improve the performance of retrieving relevant cases automatically with a query case as input [42, 44]. Existing works have indicated that the search behavior of legal case retrieval may be different from that of ordinary search (e.g., users tend to issue more queries interactively [14]). Therefore, understanding the differences may be of vital significance in the construction of practical legal case retrieval systems.

Legal case retrieval differs from general web search in various aspects. For instance, the target users in legal case retrieval are mainly legal workers with professional knowledge in law. Compared with web search engines, legal databases (e.g., Westlaw) are preferred considering the data authority and legal effect [2]. The retrieved results are mostly semi-structured case documents rather than web pages. Moreover, the definition of legal relevance [14, 45] is beyond topical relevance, involving similar legal issues, facts, consequences, etc. These differences may cause differences in user behavior and further challenge the application of user models developed for general web search in legal case retrieval.

As far as we know, user behavior in legal case retrieval is still under-investigated. Existing user-centered studies in legal domain [23, 30, 38, 50] mostly worked on the information seeking process of general legal research. However, different from legal case retrieval, it involves a range of resources and user backgrounds. The corresponding information-seeking models are sometimes presented as a high level of abstraction [30], leaving a gap from specific information retrieval tasks. Besides, previous works were mainly based on qualitative methods, such as interviews [23, 30, 47], surveys [38],

and monitoring [50], while fine-grained user behavior in the search process still lacks a thorough and quantitative investigation.

Both the user and the search task play vital roles in the interactive search process. In particular, the influence of domain expertise and task difficulty is worth investigating. From the user perspective, previous works [32, 38, 48] mainly investigated domain expertise generally based on users' majors (e.g., politics, medicine, law). While most of the users in legal case retrieval major in law, they still vary in domain expertise depending on their specialized fields, such as criminal law, civil law, and administrative law. Meanwhile, task difficulty is one of the significant factors from the search task perspective [3, 8, 27]. Especially in judicial practice, the applicable procedures may also vary according to the case difficulty [20, 51].

In this paper, we focus on an in-depth investigation of user behavior in legal case retrieval, raising the following research questions,

- **RQ1**: *How do legal practitioners conduct legal case retrieval? What are the differences from general Web search?*
- **RQ2**: *What factors affect user behavior in the legal case retrieval process?*

To shed light on the application of user behavior in this search scenario, we propose the third research question:

- **RQ3**: *How can we learn from users' implicit feedback in legal case retrieval?*

To address these research questions, we conducted a laboratory user study ($N = 45$). With the participants who majored in law but had different legal specialties, we logged rich behavioral data in the search process, including queries, click-through, hovering, scrolling, and dwell time. We also collected users' self-reported feedback and relevance assessments in the study. With the collected data, we systematically investigate users' search behavior during legal case retrieval. In summary, our key contributions are four-fold:

- We collect a behavioral dataset along with relevance assessments in legal case retrieval. The dataset[1] is open now.
- We generalize the properties of the search process in legal case retrieval and analyze its differences from general web search quantitatively based on various behavioral measures.
- We provide a thorough analysis of how task difficulty and domain expertise affect user behavior in this search scenario.
- We investigate users' implicit feedback for query efficiency and case relevance, and further propose a predictive model for relevance feedback.

2 RELATED WORK

In general web search, user behavior has drawn active research interest and benefited related IR tasks [1, 7, 9, 17]. Specifically, domain expertise and task difficulty are two influencing factors. White et al. [48] conducted a longitudinal analysis to investigate the impacts of domain expertise on web search behavior in four different domains (medicine, finance, law, and computer science). Mao et al. [32] further conducted a laboratory study to investigate the effects of domain expertise in exploratory search, involving tasks of environment, medicine, and political domains. As for task difficulty, multiple research efforts have been made to analyze the effects on search behavior in web search and to build corresponding prediction models [3, 8, 27]. We note that there is no consensus on the definition of task difficulty. Some studies [21, 25] defined difficulty as the user's subjective perception of task complexity, while some studies [3, 26] considered objective measurements, e.g., correctness, task performance.

Professional search is defined as an activity to support or address the work tasks of professionals within various domains [38, 43], e.g., patent, legal, healthcare, etc. A recent workshop [46] summarized three key characteristics of professional search as heterogeneous information sources, highly interactive procedure, and highly specialized activities. Different from general web search [6], professional search usually have domain-specific requirements [38, 46]. Russell-Rose et al. [38] conducted a comparison study of four kinds of professionals by analyzing surveys of purposive samples. Their results revealed that different professionals prioritized different features and functions, although they shared some fundamental needs. Within a specific domain, Liu & Wacholder [29] emphasized the role of domain expertise in exploiting MeSH terms by evaluating different types of searchers in medical search.

In the earlier research line of legal IR, extensive expert efforts were involved in building legal information retrieval systems (e.g., Westlaw [12]). Both knowledge engineering-based techniques [37, 39] and NLP-based methods [4, 34] were explored. Several empirical studies have been performed to study the general legal information-seeking procedure. For instance, Kuhlthau & Tama [23] conducted structured interviews with eight lawyers and framed the information-seeking process with ISP model [22]. Makri et al. [30] investigated the application of Ellis's model [13] to legal information-seeking behavior via semi-structured interviews and naturalistic observations of 27 academic lawyers. In particular, legal case retrieval is a specialized task that aims to retrieve supporting prior cases given a query case. Recent benchmarks, such as COLIEE [36] and AILA [5], aimed to explore the NLP-based methods and contributed several moderate datasets for this task. Following these benchmarks, prior efforts [42, 44] were put into developing automatic retrieval models.

However, in the practice of legal case retrieval, users tend to search interactively instead of querying with an entire case directly. Therefore, there still exists a considerable gap between the standard benchmark and the legal practice. Unlike existing research that involves various domains or information resources, our work focuses on a specific but fundamental legal search task and investigate users' fine-grained search behavior in a quantitative way.

3 USER STUDY

3.1 Tasks

We adopted the taxonomy of litigation cases (i.e., criminal, civil, administrative) that is popularly accepted in the Chinese law system [49] and designed one search task of each category. We also included an additional query case as the warm-up task to familiarize the participants with the experimental environment. An expert (Ph.D. in Law) designed the search tasks. These tasks were in different difficulty levels, among which the criminal task was the most difficult. The query cases were all adapted from the real effective judgments. Following previous work [36], the query case should remain un-judged when the legal case retrieval process is

[1]Now available at https://github.com/ThuYShao/UserStudyLegalDataset.git

Table 1: An example of the search task and the corresponding Causes.

Category	Query Case Description	Causes of Action
Criminal	The defendant Zhang took advantage of the feature that the capital accounts of the same business department in a securities company had the identical first four digits and contained the same and simple initial passwords. In order to practice the skills for speculating on the stock market, the defendant obtained the accounts and passwords of 120 customers in the telephone commission system of an electronic trading center of the company business department through continuous trial and error login using a video phone in this residence, in early January 2011. The defendant then conducted unauthorized stock trading via 10 of these client accounts. He authorized more than 200 orders and completed nearly 100 transactions. The money involved in these transactions added up to over 15,000,000 Yuan and the economic loss for the customers reached 130,000 Yuan in total.	Crime of intentional destruction of property; Crime of illegal invasion computer information system

performed, so we removed the parts containing the court's opinions. We anonymized the query cases to ensure that the original judgments could not be easily retrieved by just querying with a specific name of a person or place in the given description. Table 1 gives an example of the tasks. Specifically, the query case description would be shown to the participants in the user study, while Causes of Action, which were determined by the original judgment, functioned as the golden reference for the external assessors in Section 3.5. Similar to existing studies [11, 28, 53] in information systems, the number of tasks was limited considering user workload.

3.2 Participants

We recruited 45 participants (11 males, 34 females) via online forums and social networks. Among them, 31 were graduate students in law school, and 14 worked in law firms. They were all native Chinese speakers and qualified in legal practice[2]. They varied in legal specialties, among whom 15 majored in criminal law, 21 in civil law, 2 in administrative law, and 7 had no specialty. The *domain expertise* was determined depending on the participant's legal specialty. If a task category was consistent with her legal specialty, the session was within the domain (*in-domain*), otherwise was out of the domain (*out-domain*). If the participant had no specialty, all of her sessions were considered to be *out-domain*.

3.3 Experimental System

We developed a user study system using Django, where participants could log in and complete the entire user study procedure. As for the experimental search system, we re-directed to a commercial legal search engine[3] to simulate the real search environment. The Google Chrome browser was used to display the experimental pages. We developed a customized chrome browser extension to log search behavior on search result pages (*SERP*s) and landing pages (*LDPage*s) and record retrieved cases. Figure 1 shows examples of the SERP and LDPage. Query suggestions and advertisements were excluded. It had been confirmed ahead that the search system would not do personalization. Similar to common legal search engines, there was a "field filter" on the SERP, which could be used to select the court's level, procedural posture, etc. If the terms (conditions) were selected, they would occur in the query along with specific field symbols. As for the landing page, we injected a button for bookmarking if the case was possibly relevant, which floated at the left corner of the page.

[2]They had passed the "National Uniform Legal Profession Qualification Examination"
[3]https://ydzk.chineselaw.com/case

3.4 Procedure

Before the experiment, participants were required to watch a video that instructed the requirements and data collection procedure. They then signed the informed consent. The study began with a warm-up task. The main tasks were shown in a random order to balance the order effects [24]. Each participant spent about 1.5 hours completing the main tasks and gained $18 for involvement.

Each task can be generally divided into two parts, i.e., a searching part and an annotating part, as illustrated in Figure 1.

Query Case Reading. In each task, the participant was instructed to assume that she is dealing with the given query case and needs to find adequate relevant cases supporting the decision process. The participant could refer to this page any time during searching, so she did not need to memorize the case description.

Pre-task questionnaire. The participant filled in a pre-task questionnaire to report her perceived task difficulty, pre-search knowledge, and interest [32]. The questions were answered on a 5-point Likert-type scale (1: not at all, 5: very).

Searching and bookmarking. The participant was directed to the experimental search engine. The participant could freely interact with the search system as she usually did, e.g., issuing queries, clicking on results, scrolling up and down, reformulating queries, etc. As an additional requirement in our user study, the participant was asked to bookmark the cases that she felt probably relevant by clicking on the button ("Mark this case") on the landing page. No restrictions on the search time or the number of bookmarked cases were imposed. Once the participant felt that she had found enough information or could not find more useful results, she could end the search session by closing the search pages and landing pages. The browser extension would record all of the visited pages and log the user's behavior at this stage, including query formulation, click-through, hover, scroll, and timestamps.

Post-task questionnaire. The participant was directed to a post-task questionnaire to report her perceived search success and search satisfaction. Similarly, all of the questions were answered on a 5-point scale.

Bookmarked case annotating. The cases were shown in order of their bookmarked timestamps. Only one case was shown each time. For each case, the participant made a relevance assessment on a 4-point scale (1: irrelevant, 2: slightly relevant, 3: fairly relevant, 4: highly relevant). The cases that were clicked but not bookmarked would be treated as irrelevant in our study. Furthermore, the participant was required to provide the reasons for her relevance assessment in a text input box. The collected reasons were used to ensure the quality of her assessments in this paper.

Figure 1: The procedure of each task and examples of the SERP and landing page in the experimental system.

After the participant finished this stage, the corresponding task was completed, and she could start another with the same procedure.

A pilot study, which involved four additional users, was conducted ahead to ensure the designed tasks, the experimental system, and the user study procedure work well.

3.5 Data Cleansing and Assessment

The sessions were filtered out if we found that the participant had known the original case before searching. We identified three such sessions via the participants' self-reports and inspecting their queries. There were another four sessions that had problems in behavior logging due to network instability. In total, we excluded seven invalid search sessions along with their retrieved cases. After filtering, we collected 128 sessions ("in-domain": 36, "out-domain": 92), 1,682 queries, and 1,289 landing pages (558 were bookmarked).

By manually inspecting some submitted queries, we identified two categories of legal concepts, i.e., the Cause of Action (*Cause*) and the Facts of Case (*Fact*). The *Fact* is usually from the detailed circumstances of a case directly, while the *Cause* is the legal generalization of the issue, which requires domain-specific knowledge for refinement. Table 1 gives examples of Causes in the given task. To investigate query formulation behavior in legal case retrieval, we recruited three external legal experts to assess the existence and the correctness of Causes in the queries that the participants submitted. In detail, as for each query term, they annotated whether it was a Cause or a Fact and whether it was correct if it was a Cause. We provided them with the courts' decisions and holdings of each query case as the golden reference for assessing Cause correctness. The Fleiss's κ of the existence and correctness assessments (both in binary scale) among three assessors were 0.9478 and 0.9246, respectively, indicating almost perfect agreement [15]. If there were disagreements, we took the result of the majority vote.

4 SEARCH BEHAVIOR ANALYSIS

4.1 Comparison with Web Search

To understand the characteristics of users' search behavior in legal case retrieval, we compare it with that in general web search.

4.1.1 Datasets for comparison. We utilize two public search behavior datasets in web search from a laboratory user study [33] (denoted as *WebUserS*) and a field study [52] (denoted as *WebFieldS*) to compare with that collected in our user study (denoted as *LegalUserS*).

Table 2: Properties of the datasets of legal case retrieval and web search.

Property	LegalUserS	WebUserS [33]	WebFieldS [52]
Source	Lab user study	Lab user study	Field study
User domain	Legal	Non-specific	Non-specific
Language	Chinese	Chinese	Chinese
# users	45	25	30
# tasks	3	9	nan.
# sessions	128	225	1,124
# queries	1,682	935	3,535
Logged behavior	Query, click, hover, mouse, timestamps	Query, click, timestamps	Query, click, mouse, timestamps

Table 2 shows the properties of these datasets. Both *WebUserS* and *WebFieldS* are in Chinese and published in recent years (2016 and 2020, respectively). The user study [33] was designed to simulate a real web search environment, in which the experimental search engine could access the open Web and support query reformulation and pagination. The field study [52] collected users' daily search activities in web search, overcoming the limitations of lab studies (e.g., the discrepancy from real search scenarios) and large-scale log analysis (e.g., noise). Both datasets included rich behavioral features. Therefore, we think that they can represent users' search process in general web search and are suitable for comparison.

4.1.2 Comparison of Behavioral Measures. We investigate search behavior from multiple aspects, including search effort within a session, query formulation, session-level examination patterns, and examination on a SERP. The median and the quartiles of each measure are shown in Table 3 to provide an overview of the data distribution and alleviate the effects of outliers (especially in the field study dataset). All p-values are calibrated through Bonferroni correction [41] within the corresponding group to deal with the multiple comparison problem [16]. Mann-Whitney U test [31] instead of the t-test is conducted since most of the variables have a non-normal distribution. We acknowledge that there are some differences between *WebUserS* [33] and *WebFieldS* [52]. But we mainly focus on investigating the difference between legal case retrieval and web search, so the statistical test is conducted between *LegalUserS* and *WebUserS* / *WebFieldS*, respectively. Results are shown in Table 3.

Table 3: Behavioral measures in legal/web datasets."[Q1, Q3]" denotes the interval composed of the upper and lower quartiles.

Group	Behavioral Measure	LegalUserS		WebUserS [33]			WebFieldS [52]		
		Median	[Q1, Q3]	Median	[Q1, Q3]	Sig.	Median	[Q1, Q3]	Sig.
Search effort within a session	task time (s)	**717.3**	[431.5, 1075]	nan.	nan.	–	94.36	[33.57, 315.1]	***
	# queries	**7.000**	[4.000, 16.25]	4.000	[2.000, 5.000]	***	2.000	[1.000, 3.000]	***
	# clicks	**8.000**	[5.000, 13.00]	6.000	[4.000, 8.000]	***	2.000	[1.000, 4.000]	***
	# pages	**18.50**	[10.75, 31.25]	10.00	[7.000, 15.00]	***	4.000	[2.000, 7.000]	***
	% query w/o click	**0.5000**	[0.1384, 0.6667]	0	[0, 0.3333]	***	0	[0, 0.2837]	***
Query formulation	avg. query length *per query*	**8.000**	[6.000, 10.00]	7.000	[5.000, 9.000]	***	6.000	[4.000, 9.000]	***
	% generalization	**0.3030**	[0.0955, 0.3957]	0	[0, 0.1667]	***	0	[0, 0]	***
	% specification	**0.4545**	[0.3333, 0.5670]	0.4000	[0.2222, 0.5000]	–	0	[0, 0.2500]	***
	% substitution	**0.2000**	[0, 0.4000]	0.5000	[0, 0.6667]	***	1.000	[0.6000, 1.000]	***
Session-level examination patterns	% time spent on SERPs	**0.3764**	[0.2856, 0.5299]	nan.	nan.	–	0.2274	[0.0766, 0.5498]	***
	% query of 1st click	0.3261	[0.1818, 0.5000]	0.3333	[0.2000, 0.5000]	–	0.3333	[0.2000, 0.5000]	–
	% query of most clicks	**0.7500**	[0.4000, 1.000]	0.5000	[0.3333, 0.8000]	***	0.5000	[0.3333, 1.000]	**
SERP examination	P(click) *per query*	**0.2000**	[0.1000, 0.4375]	0.1000	[0.1000, 0.2000]	***	0.1000	[0.1000, 0.2000]	***
	avg. click rank *per query*	2.500	[1.500, 4.500]	3.225	[2.000, 5.000]	***	2.000	[1.000, 3.000]	***
	max click rank *per query*	3.000	[2.000, 7.000]	4.000	[2.000, 8.000]	*	2.000	[1.000, 4.000]	***
	time (s) to first click *per query*	**13.47**	[8.245, 24.37]	6.381	[3.875, 9.872]	***	5.325	[3.250, 9.853]	***
	avg. click dwell time (s) *per query*	**38.43**	[23.81, 63.66]	26.95	[12.42, 51.46]	***	19.62	[8.579, 56.60]	***
	# avg. skipped results between clicks *per query*	**0.6667**	[0, 1.667]	1.667	[0.6667, 3.000]	***	1.000	[0, 2.000]	***

"*/**/***" indicates the difference in the measure is statistically significant at $p < 0.05/0.01/0.001$ (after Bonferroni correction) level compared with *LegalUserS*. "nan." indicates the measure is unavailable in the dataset.

Search Effort within a Session. Legal case retrieval involves a much higher search effort compared with web search. On the one hand, the submitted queries, clicked results, visited pages, and the total task time increase significantly in legal case retrieval. On the other hand, a larger proportion of the query lacking clicks suggests that the submitted queries are more likely to fail in satisfying the user's information need [10, 19]. Given that, the legal case retrieval process also requires more effort to formulate queries.

Query Formulation. We first inspect the behavioral measures that can be calculated identically in all datasets. To be specific, the *query length* is calculated based on Chinese characters. The query reformulation type is automatically determined following the previous work [32]. From Table 3, users appear to formulate longer queries in legal case retrieval. Compared with web search, *generalization* and *specification* make up larger proportions, indicating that the query reformulation involves multiple careful trials.

Moreover, while keyword-based queries are widely adopted in current web search systems, multiple querying methods are employed in legal case retrieval. For instance, besides keyword-based query terms, 76 of 128 sessions in our user study involve conditions from the "field filter". The *Cause of Action*, legal abstraction of the case issue, accounts for a considerable part of the queries in legal case retrieval (e.g., 95 sessions involve Causes).

Session-level Examination Patterns. Among its measures in Table 3, *% query of 1st / most click* represents the normalized position of the query that involves the first / most clicks within a session. We treat them as two representative statistics of the distribution of clicks within the search session. Moreover, Figure 2 gives an intuitive view of the click patterns within a session. The distribution in legal case retrieval is significantly different from that in web search (K-S two-sample test, $p < 0.001$, respectively), while no significant difference is detected between that of two web datasets ($p = 0.12$). Unlike web search sessions that involve multiple peaks of click, clicks concentrate on the latter part of the session in legal case retrieval. In particular, the query involving the most clicks occurs later. Besides, the difference in *% time spent on SERPs* also indicates

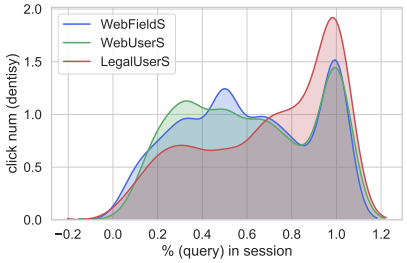

Figure 2: Distribution of the number of clicks within a session in three datasets.

users pay more attention to exploring the SERPs in legal case retrieval. Therefore, the search process of legal case retrieval appears to be more exploratory, involving multiple trials, modification, and learning before an in-depth examination.

SERP Examination. To be specific, *time to first click* denotes the time interval from displaying the results to the user's first click, and *# avg. skipped results between clicks* means the average number of skipped results in the click sequence. These measures represent users' patience, carefulness, and decision speed when examining the SERP. We can observe several consistent trends when comparing to both web datasets. In legal case retrieval, users show slower decision speed since it takes longer before their first click, and they spend more time examining the clicked results. They also appear to be more patient when examining the result lists according to the click rate and the number of skipped results.

4.1.3 Summary. Regarding **RQ1**, the legal case retrieval process can be characterized as an exploratory search process, consistent with previous work in professional search [46]. Compared with general web search, users pay more search effort and seem more persistent in satisfying their information needs. They appear to be

more patient and careful when reformulating queries and examining results. Besides, domain-specific knowledge is incorporated into the search process, e.g., refining Causes in query formulation.

4.2 Effects of Task Difficulty and Domain Expertise

To address **RQ2**, we investigate the search process of legal case retrieval in-depth. We focus on inspecting two independent variables, i.e., *domain expertise* and *task difficulty*. As mentioned in Section 3, search sessions are classified into *in-domain* and *out-domain* groups based on the user's legal specialty. The designed task difficulty is verified by users' perceptions collected in the pre-task questionnaires. To be specific, the user-reported difficulty level of the criminal task is significantly higher than the others, while the difference between the civil and the administrative is insignificant (*avg. difficulty*, 3.277 v.s. 2.756 / 2.535, by Dunn's test, $p < 0.05$ after B-H adjustment). Therefore, we group them into two difficulty levels, taking the criminal task as the *difficult* (*hard*) while the civil and administrative tasks as the *easy*.

4.2.1 Behavioral Measures. As shown in Table 4, we investigate the effects of task difficulty and domain expertise on the behavioral measures of various categories. As most of the measures are not normally distributed, we use non-parametric statistical tests. We first conduct the Scheirer Ray Hare Test [40], a non-parametric test for a two-way factorial experiment, using task difficulty and domain expertise as two factors. Results are given in Table 4. Generally, task difficulty has a significant impact on a larger proportion of behavioral measures than domain expertise. The Mann-Whitney U test is further conducted within each domain expertise to examine the effects of task difficulty under different expertise.

Search Effort within a Session. Besides the measures used in Section 4.1, we inspect the mouse-hovering behavior. Mouse-hovering is also an indicator of examination, which correlates with user satisfaction but involves less examination workload than click-through [10]. As shown in Table 4, search effort increases significantly in difficult settings, reflected by the increase of issued queries, visited pages, hovered results, and queries lacking interactions. Moreover, task difficulty affects search effort of *out-domain* users significantly while has few influence within the *in-domain* group. Note that both task difficulty and domain expertise do not significantly affect the number of clicks and task time. One possible explanation is that users might invest limited patient and effort in the laboratory study environment.

Query Formulation. We investigate query formulation from two perspectives, i.e., term-based and Cause-based. Regarding term-based behavior, we find that task difficulty has significant influences. Furthermore, task difficulty impacts the corresponding measures in both domain expertise groups. Users explore more various query terms in difficult tasks. In contrast, a larger proportion of *specification* in easy tasks indicates that users might identify a specific search direction and exploit it continuously.

Different from *Fact*, *Cause* is a high-level summary of legal issues in a case. Based on the external annotations, we inspect the usage of Cause in users' query formulation process. Consistent with the results of term-based analysis, the use of Cause is mainly influenced by task difficulty. Users are more likely to identify Cause precisely in easy tasks while the accuracy drops significantly encountering difficult settings. Besides, the Cause is used later in the sessions of difficult tasks, and it takes more trials before users could identify the correct one. The differences suggest a more exploratory and struggling query formulation process in difficult tasks, in which users constantly locate and revise the core legal relationships of the query case. The effects are consistent in both expertise groups and more significant in the *out-domain* group.

Session-level Examination Patterns. Users spend a larger proportion of time on SERPs in difficult tasks. Both *in-domain* and *out-domain* users are affected by task difficulty while the impact is greater on *out-domain* users. However, we do not observe any significant influence of task difficulty or domain expertise on the click patterns within a session.

SERP Examination. In particular, we consider hover and click as examination behavior of different levels. To be specific, users usually hover on a result item to make a preliminary judgment before clicking for further examination. In Table 4, *P(click|hover)* denotes the probability of a result to be clicked given hovered. We observe that task difficulty influences this measure significantly, consistent in both domain expertise groups. It suggests that users put more effort into identifying relevant items on SERPs and might skip more results based on the preliminary judgment when task difficulty increases. Besides, we find that users are more likely to click on the results ranked low in easy tasks. Since users are more likely to formulate efficient queries in easy settings, the returned result list should be more relevant. Thus, users might explore deeper of the result list. No significant effects are observed on other measures.

LDPage Examination. Unlike web search results, the landing pages of legal case retrieval are mostly case documents, which always contain long texts, so the screen can hardly display the entire document one time. Moreover, a case document is semi-structured in content, containing multiple sections. Table 5 gives an example in which the section ID denotes its position in the document. Users need to scroll to examine different sections, causing changes in viewport. The viewport means the portion of the document page that is visible on the screen at a certain time. In this paper, we use the viewport to simulate users' examination attention.

Previous works [24, 54] found a strong correlation between viewport with user attention. We utilized the *weighted* viewport to reduce the presentation bias, which had the strongest correlation with user attention [24, 54]. The weighting factor is calculated following $\omega_s^i = \left(h_{v,s}^i\right)^2 / \left(h_v^i * h_s\right)$, where $h_{v,s}^i$ is the visible height of section s in the i-th viewport, h_v^i is the height of the i-th viewport, and h_s is the actual height of section s. In table 4, *pos. of first viewport* is the weighted section IDs in the first viewport. *total viewport time* is the sum of the weighted viewport time of all sections in all viewports. We view the consecutive viewports that contain the same sections as *stationary* (*merged* viewport). Then the *moved up (down) viewport* denotes that a section with a smaller (larger) ID occurs in the $(i+1)$-th viewport. *# skipped sections* is the number of sections that do not occur in any viewports. *# revisit* means the sum of times a section is revisited by a non-stationary viewport.

From Table 4, users appear to examine the case document mainly in a top-down manner (*% moved down* > *% moved up*), while some sections are revisited multiple times. In difficult tasks, users are

Table 4: Differences in behavioral measures w.r.t different task difficulty and domain expertise.

Group	Behavioral Measure	Easy	Hard	In-d	Out-d	In-d			Out-d		
						Easy	Hard	Sig.	Easy	Hard	Sig.
Search effort within a session	# queries †††	7.131	**20.61**	12.00	11.67	8.048	17.53	–	6.825	**22.21**	***
	# pages †††	17.86	**29.52**	22.00	21.82	19.24	25.87	–	17.40	**31.41**	***
	# clicks	10.73	8.909	10.00	10.14	11.19	8.333	–	10.57	9.207	–
	# hovers †††	47.26	**94.45**	69.92	60.97	57.24	87.67	–	43.94	**97.97**	***
	task time (s)	785.6	904.1	885.2	803.3	916.7	841.3	–	742.0	936.7	–
	% query w/o click †††	0.3523	**0.5873**	0.4452	0.4286	0.4439	0.4470	–	0.3222	**0.6598**	***
	% query w/o hover ††	0.1933	**0.3189**	0.2401	0.2350	0.1860	0.3160	–	0.1957	**0.3204**	**
Query formulation (term-based)	# unique terms *per session* †††	5.560	**11.50**	7.694	7.565	5.762	10.4	–	5.492	**12.07**	***
	% generalization	0.2358	0.3216	0.2102	0.2923	0.2164	0.2015	–	0.2433	**0.3837**	**
	% specification †††	**0.5249**	0.3431	0.4813	0.4475	**0.6422**	0.2561	***	0.4793	0.3881	–
	% substitution †	0.2393	**0.3343**	0.3085	0.2602	0.1414	**0.5425**	**	0.2773	0.2282	–
Query formulation (Cause-based)	# unique Causes *per session*	2.000	2.467	1.643	2.358	1.500	1.833	–	2.163	2.889	–
	% correct Cause *per session* †††	**0.9405**	0.4515	0.7768	0.7900	**1.000**	0.4792	***	**0.9211**	0.4330	***
	query (pos. in session) w/ 1st Cause †††	2.015	**7.700**	4.250	3.627	1.688	7.667	–	2.122	**7.722**	***
	distance (absolute value) between query w/ 1st Cause and query w/ 1st correct Cause †††	0.2308	**7.967**	1.250	3.269	0	**2.917**	**	0.3061	**11.33**	***
	above-mentioned distance normalized by session length †††	0.0346	**0.3896**	0.1176	0.1589	0	**0.2744**	**	0.0459	**0.4665**	***
Session-level examination patterns	% time spent on SERPs †††	0.3368	**0.5887**	0.4333	0.4195	0.3403	**0.5635**	**	0.3356	**0.6017**	***
	% query of 1st click	0.4076	0.3796	0.4631	0.3687	0.4473	0.4853	–	0.3922	0.3249	–
	% query of most clicks	0.6743	0.7237	0.7230	0.6794	0.6315	0.8509	–	0.6909	0.6580	–
SERP examination	P(click) *per query*	0.3389	0.3153	0.3480	0.3224	0.346	0.3509	–	0.3362	0.3007	–
	P(hover) *per query*	0.9203	0.7835	0.8241	0.8549	0.9149	0.7514	–	0.9226	0.7980	–
	P(click|hover) *per query* †††	**0.2431**	0.1115	0.1629	0.1747	**0.2392**	0.1018	**	**0.2448**	0.1158	***
	avg.click rank *per query*	4.138	3.642	3.943	3.943	4.141	3.652	–	4.137	3.638	–
	max. click rank *per query* †	**6.607**	5.377	6.246	6.075	6.747	5.510	–	**6.554**	5.323	*
	time (s) to first click *per query*	17.63	22.75	16.50	20.89	12.06	23.03	–	19.78	22.64	–
	avg. click dwell time (s) *per query*	51.28	42.71	50.22	47.00	**56.14**	41.52	*	49.41	43.20	–
	# avg. skipped results between clicks *per query*	1.249	1.701	1.242	1.903	1.268	1.208	–	1.242	1.903	–
LDPage examination	pos. of first viewport *per page*	0.1602	0.1415	0.1481	0.1569	0.1551	0.1352	–	0.162	0.1444	–
	total viewport time (s) *per page*	16.16	15.45	16.04	15.91	17.05	14.15	–	15.85	16.05	–
	% moved up viewport *per page*	0.2156	0.2034	0.2087	0.2132	0.2128	0.2009	–	0.2166	0.2045	–
	% moved down viewport *per page* ††	0.4146	**0.4461**	0.4221	0.4249	0.4003	**0.4626**	**	0.4196	0.4383	–
	# skipped sections *per page* ††† ‡ ‡‡	0.9076	**1.217**	0.7905	**1.083**	0.6137	**1.120**	***	1.011	**1.263**	*
	# revisit *per page*	10.22	9.744	9.964	10.12	10.733	8.528	–	10.04	10.31	–

"†/††/†††" ("‡/‡ ‡ /‡ ‡ ‡") indicates that task difficulty (domain expertise) has a significant effect on the measure at $p < 0.05/0.01/0.001$ level by Scheirer Ray Hare Test (after Bonferroni correction). "*/**/***" indicates that the effects of task difficulty on the measure within one domain expertise group is significant at $p < 0.05/0.01/0.001$ (after Bonferroni correction) level using Mann-Whitney U test.

Table 5: An example of the case document

Section (ID)	Content (part)
Party Information (1)	Defendant: Yang, male, born in [Place] on [Date], of the XX nationality, with [education], unemployed...
Procedural Posture (2)	The Intermediate People's Court, after trying the case on public prosecution by the People's procuratorate against Yang on the crime of trafficking in drugs, ascertained on [Date] by Criminal Judgment No. 52 (2000)...
Facts (3)	One day in July 1998, under the arrangement of Ren XX, and via Yi XX, Yang sold to Cao, a drugger, in Ren's "Aliang Hair Care Salon", 100 grams of heroin offered by Ren, and got 19,000 Yuan of illicit money...
Holdings (4)	This Court holds that Yang's offences of illegally trafficking in and transporting heroin have constituted the crimes of trafficking in and transporting drugs, and the quantity was large, thus he should be punished in accordance with the law...
Decision (5)	I.the part on sentencing the punishments on Yang in the Criminal Judgment No. 84 (2001) in the Final Instance by the Criminal Division of XX Higher Court and in the Criminal Judgment No. 52 (2000) shall be rescinded...
End of Document (6)	Chief Judge XXX; ...; Date XXX; Court Clerk XXX

more likely to read the case in a top-down sequence but skip more sections. Meanwhile, *out-domain* users appear to skip sections more. We do not observe any significant influence of domain expertise or task difficulty on other viewport measures.

4.2.2 Summary. Regarding **RQ2**, task difficulty has a more significant impact on search behavior than legal-specific domain expertise. Particularly, task difficulty affects query formulation behavior of different domain expertise, while the effects on search effort mainly occur in *out-domain* users. Compared with the session-level behavioral measures, those measuring examinations on a specific page (SERP or LDPage) are less affected.

5 IMPLICIT FEEDBACK

To address **RQ3**, we look into users' implicit feedback from two perspectives. On the one hand, we inspect the efficiency of query formulation by taking click-through and bookmarking as implicit feedback. On the other, we investigate whether viewport behavior on the landing page correlates with user-annotated relevance and how it contributes to relevance feedback.

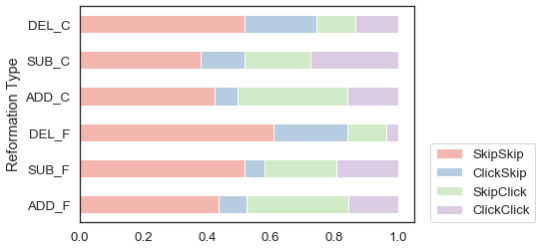

Figure 3: Proportion of click patterns (SkipSkip / ClickSkip / SkipClick / ClickClick) under each reformulation type.

Table 6: Differences in implicit feedback w.r.t. queries.

Implicit Feedback	w/o C	w/o TC	w/ TC
P(click) *per query* ***	0.0943	0.0413 †	**0.2040** †‡
P(bookmark) *per query* ***	0.0376	0.0142 †	**0.1068** †‡

"***" indicates the difference is significant at $p < 0.001$ using Kruskal-Wallis Test. †/‡ indicates the result is significant different from that of "w/o C"/"w/o TC" at $p < 0.05$ by Dunn's Test (after correction).

5.1 Implicit Feedback for Query Efficiency

First, we look into the query reformulation efficiency with click-through as implicit feedback. Following previous work [19], we inspect four possible click patterns. In detail, *SkipClick* means that the user does not click any result in the i-th query, then reformulates the query and clicks the results in the $(i+1)$-th query. It indicates that the query reformulation is effective, while the *ClickSkip* pattern suggests that the reformulation does not help. Similarly, the consecutive clicks or skips (i.e., *ClickClick* or *SkipSkip*) can be viewed as an indicator of successful (or failed) searches.

Beyond the reformulation type classified based on terms, we inspect the modification of *Cause* and *Fact* in a query. The query reformulation type is defined based on *deleting* / *adding* / *substituting* the *Cause* / *Fact* terms. Figure 3 shows the proportion of the click patterns. The reformulation type has a significant influence on the click pattern ($\chi^2_{15}, p < 0.001$). Looking at the ratio of *SkipSkip* + *ClickSkip*, we find the deletion of Cause or Fact indicates lower efficiency than the other types. In particular, deleting Fact terms shows a higher proportion of *SkipSkip*, indicating failed searches. When the user deletes a query term, she might not find proper querying directions, so the reformulation efficiency is lower. Meanwhile, we observe adding new terms (Fact or Cause) contributes to a higher proportion of *SkipClick* pattern, indicating that users might be unsatisfied with the initial query but the reformulation helps users to find relevant results. We think that this is due to users might identify some useful query conditions during the reformulation process. In general, similar trends can be observed in modifying Cause and Fact terms.

We further investigate the role of *Cause* in each query with click-through and bookmarking as implicit feedback. In Table 6, a higher value in each feedback is considered as an indicator of higher query efficiency. Specifically, bookmarking indicates the usefulness of a clicked result. Queries are classified into three groups, i.e., the query without Cause (*w/o C*), the query including Causes but without correct Cause (*w/o TC*), and the query with correct Cause (*w/ TC*). We assume that users might not understand the key issues of a query case well if they can not identify a correct Cause, and the corresponding query would be less helpful. As a result, the existence and the correctness of Cause in a query significantly impact the two measures. The query including a correct Cause can contribute to the highest click-through and bookmarking rate, which is the most efficient. In contrast, the lack of Cause or the incorrect Cause leads to a non-trivial drop. The results emphasize the vital role of a correct Cause in query formulation.

5.2 Implicit Feedback for Relevance

Inspired by the application of viewport in mobile search [24, 54], we inspect how it correlates with relevance in legal case retrieval. Besides the viewport features calculated on the entire landing page, we further inspect those of the main sections. As shown in Table 7, we calculate five viewport features for each section. Among them, *time to first viewport* means the time interval from displaying the landing page to the first appearance of the section in a viewport. *max. viewport time* denotes the section's maximum weighted time in a merged viewport. *time of last viewport* denotes the weighted time in the section's last merged viewport.

5.2.1 Correlation Analysis. Table 7 presents the Spearman's rank correlation between each viewport feature and relevance. Among the features calculated on the entire page (*General*), the total viewport time and the number of revisits have significant positive correlations with relevance, while the number of skipped sections has a significant negative correlation. The longer viewport time spent on a case document, more revisits, or fewer skips indicate the higher relevance. Meanwhile, viewport features of certain sections (i.e., Facts, Holdings, Decision, and End of Document) correlate with relevance significantly. In the legal domain, Facts, Holdings, and Decision are usually considered as the core parts of a case document. As shown in Table 7, users tend to spend more time on these sections, read slower, and revisit them for more times if the document is relevant. Especially for the Decision section, we observe that users might examine it for a longer time before leaving if relevant. End of Document, as the last part of a document, mainly contains meta-information, e.g., the collegial panel members, the date, etc. We explain the correlation by that if the case is relevant, a user might read through the document patiently till the *EoD*.

5.2.2 Prediction. We further apply these features to relevance prediction and inspect their roles for relevance feedback. Each group of viewport features are defined as above, and we combine the features of *General* and the last four sections (i.e., *F*, *H*, *D*, and *EoD*) since they have some significant correlations with relevance. To evaluate the effectiveness, we use textual features and a random classifier as the baselines. The textual features are constructed based on the standard BM25 scores (implemented by Gensim). Beyond matching the query case with the whole document, we also calculate the BM25 score based on matching each section. We treat relevance prediction as a binary classification problem. In detail, the cases with scores of 1 & 2 are considered irrelevant, and the others are relevant (irrelevant: 855, relevant: 434). The prediction is conducted on stratified 10-fold cross-validation and leave-one-user-out validation. Due to the imbalanced distribution of labels,

Table 7: Spearman's ρ between the viewport features and case relevance (4-level).

Viewport Feature	General	Viewport Feature	Sections					
			PI	PP	F	H	D	EoD
pos. of first viewport	-0.0064	time to first viewport	0.0813	0.0768*	0.0301	0.2394***	0.2207***	0.2864***
total viewport time	0.2499***	total viewport time	0.0160	0.0107	0.1988***	0.1303***	0.2211***	0.0933*
% move up viewport	0.0227	max. viewport time	0.0154	0.0015	0.1635***	0.1291***	0.2212***	0.0962*
% move down viewport	-0.0522	time of last viewport	-0.0169	-0.0260	-0.0726*	0.0538	0.1895***	0.1058***
# revisit sections	0.2096***	# revisit	0.0368	0.0513	0.2282***	0.1154***	0.1769***	0.0475
# skipped sections	-0.1387***	–	–	–	–	–	–	–

PI/PP/F/H/D/EoD denotes "Party Information"/"Procedural Posture"/"Facts"/"Holdings"/"Decision"/"End of Document" section in a case document. "*/**/***" indicates the correlation is significant at $p < 0.05/0.01/0.001$ level (after correction).

Table 8: Performance of relevance prediction.

Feature Group		AUC (10-fold)	AUC (User)
Viewport	General (6)	0.6415	0.6348
	Party Information (5)	0.5809	0.5659
	Procedural Posture (5)	0.5616	0.5502
	Facts (5)	0.6429	0.6355
	Holdings (5)	0.6601	0.6622
	Decision (5)	0.6635	0.6683
	End of Document (5)	0.6683	0.6525
	Combined (26)	**0.7003**	**0.6995**
Text	Document (1)	0.6585	0.6361
	Sections (6)	0.6882	0.6859
	Combined (7)	0.6912	0.6853
Viewport & Text (33)		**0.7453**	**0.7402**
Random Classifier		0.4916	0.4892

The number in parentheses denotes the dimension of each feature group. "10-fold / User" denotes 10-fold cross-validation and leave-one-user-out validation.

we evaluate with AUC. As for the supervised learning methods, we have trained various classifiers, including logistic regression, decision tree, random forest, and GBDT. They reveal similar findings when comparing among the different feature groups. Among them, the random forest achieves the best prediction performance. Therefore, we mainly analyze the results given by random forest (implemented by sklearn).

Table 8 reports the prediction performances of corresponding feature groups. The results verify the effectiveness of viewport features in relevance feedback. For instance, the viewport features (*Combined*) can achieve a significant better performance than the query-document BM25 (t-test, $p < 0.05$, same as below). The viewport features and textual features are also complementary. Combination of them (*Viewport & Text*) can outperform each category.

Moreover, utilizing the document structure can benefit relevance prediction. Compared with viewport features calculated on the entire document (*General*), combining with those of the last four sections (*Combined*) can improve the performance significantly. Similarly, as for the textual features, involving the feature of each section can outperform query-document matching. In particular, viewport features of different sections achieve different performances. The prediction results are consistent with those of correlation analysis. Compared with features of Facts, those of Holdings and Decision have stronger correlations with relevance and perform better in prediction. Since Holdings and Decision are the judge's final opinions, they support generalizing the key issues of the case and therefore are vital for relevance feedback. The Facts section, which describes the case circumstances, is also valuable in relevance estimation while less critical. It also inspires us to think about the definition of legal relevance, for which the similarity in key issues rather than in detailed circumstances should be considered primarily.

6 CONCLUSION AND FUTURE WORK

In this paper, we focus on investigating user behavior in legal case retrieval. Centered on three research questions, we conducted a user study and have obtained several interesting findings. (1) Legal case retrieval can be characterized as an exploratory search process. Compared with general web search, users put more search effort, incorporate domain-specific knowledge, and appear more patient and careful during the search. (2) Task difficulty has a greater influence on search behavior than domain expertise, especially on session-level behavior. The effects on search effort and query formulation (*Cause-based*) are more significant for *out-domain* users. (3) According to implicit feedback, utilizing a correct Cause in query formulation contributes to higher efficiency while deleting Cause or Fact terms indicates the inefficiency of query reformulation. Viewport behavior on the landing page correlates with relevance and can be further applied to relevance feedback. Particularly, specific sections (e.g., Holdings and Decision) that indicate key issues should be emphasized regarding legal case relevance.

Our results have promising implications for the design of related features in legal search systems, such as task difficulty prediction, query suggestion, and relevance feedback. Our implications are not limited to legal case retrieval but also shed light on other similar search scenarios, e.g., patent retrieval and literature search.

We acknowledge some potential limitations of our study. First, the laboratory user study still varies from the real search scenario in some way. Second, the number of users and tasks is limited as in most user studies, especially those involving domain knowledge or complex tasks [11, 28, 53]. Third, the document structure inspected in this paper is typical in the Chinese law system, which might need to be retrained or adjusted depending on different law systems.

As for future work, a large-scale log analysis or a field study is promising. Moreover, a further understanding of legal IR relevance is worth investigating, which will support a series of related tasks, such as constructing large-scale datasets, enhancing document ranking, and providing better query suggestions.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (2018YFC0831700), Natural Science Foundation of China (Grant No. 61732008, 61532011, 62002194), Beijing Academy of Artificial Intelligence (BAAI), and Tsinghua University Guoqiang Research Institute.

REFERENCES

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 19–26.

[2] Olufunmilayo B Arewa. 2006. Open access in a closed universe: Lexis, Westlaw, law schools, and the legal information market. *Lewis & Clark L. Rev.* 10 (2006), 797.

[3] Anne Aula, Rehan M Khan, and Zhiwei Guan. 2010. How does search behavior change as search becomes more difficult?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 35–44.

[4] Trevor Bench-Capon, Michał Araszkiewicz, Kevin Ashley, Katie Atkinson, Floris Bex, Filipe Borges, Daniele Bourcier, Paul Bourgine, Jack G Conrad, Enrico Francesconi, et al. 2012. A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. *Artificial Intelligence and Law* 20, 3 (2012), 215–319.

[5] Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019. Overview of the FIRE 2019 AILA Track: Artificial Intelligence for Legal Assistance.. In *FIRE (Working Notes)*. 1–12.

[6] Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM New York, NY, USA, 3–10.

[7] Marc Bron, Jasmijn Van Gorp, Frank Nack, Lotte Belice Baltussen, and Maarten de Rijke. 2013. Aggregated search interface preferences in multi-session search tasks. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 123–132.

[8] Katriina Byström and Kalervo Järvelin. 1995. Task complexity affects information seeking and use. *Information processing & management* 31, 2 (1995), 191–213.

[9] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. 2008. Context-aware query suggestion by mining click-through and session data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 875–883.

[10] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 15–24.

[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 *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. 163–172.

[12] John Doyle. 1992. WESTLAW and the American digest classification scheme. *Law. Libr. J.* 84 (1992), 229.

[13] David Ellis, Deborah Cox, and Katherine Hall. 1993. A comparison of the information seeking patterns of researchers in the physical and social sciences. *Journal of documentation* (1993).

[14] Ángel Sancho Ferrer, Carlos Fernández Hernández, and Pierre Boulat. [n.d.]. LEGAL SEARCH: foundations, evolution and next challenges. The Wolters Kluwer experience LA BÚSQUEDA DE INFORMACIÓN LEGAL: fundamentos, evolución y próximos desafíos. La Experiencia de Wolters Kluwer. ([n. d.]).

[15] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[16] Norbert Fuhr. 2018. Some common mistakes in IR evaluation, and how they can be avoided. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 32–41.

[17] Qi Guo and Eugene Agichtein. 2012. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *Proceedings of the 21st international conference on World Wide Web*. 569–578.

[18] Hanjo Hamann. 2019. The German Federal Courts Dataset 1950–2019: From Paper Archives to Linked Open Data. *Journal of Empirical Legal Studies* 16, 3 (2019), 671–688.

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

[20] Jörg-Martin Jehle, Marianne Wade, and Beatrix Elsner. 2008. Prosecution and diversion within criminal justice systems in Europe. Aims and design of a comparative study. *European journal on criminal policy and research* 14, 2-3 (2008), 93–99.

[21] Jeonghyun Kim. 2006. Task difficulty as a predictor and indicator of web searching interaction. In *CHI'06 extended abstracts on human factors in computing systems*. 959–964.

[22] Carol C Kuhlthau. 1993. A principle of uncertainty for information seeking. *Journal of documentation* (1993).

[23] Carol Collier Kuhlthau and Stephanie L Tama. 2001. Information search process of lawyers: a call for'just for me'information services. *Journal of documentation* (2001).

[24] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 113–122.

[25] Yuelin Li. 2008. *Relationships among work tasks, search tasks, and interactive information searching behavior*. Ph.D. Dissertation. Rutgers University-Graduate School-New Brunswick.

[26] Jingjing Liu, Jacek Gwizdka, Chang Liu, and Nicholas J Belkin. 2010. Predicting task difficulty for different task types. *Proceedings of the American Society for Information Science and Technology* 47, 1 (2010), 1–10.

[27] Jingjing Liu, Chang Liu, Michael Cole, Nicholas J Belkin, and Xiangmin Zhang. 2012. Exploring and predicting search task difficulty. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. 1313–1322.

[28] Jiqun Liu, Matthew Mitsui, Nicholas J Belkin, and Chirag Shah. 2019. Task, information seeking intentions, and user behavior: Toward a multi-level understanding of Web search. In *Proceedings of the 2019 conference on human information interaction and retrieval*. 123–132.

[29] Ying-Hsang Liu and Nina Wacholder. 2017. Evaluating the impact of MeSH (Medical Subject Headings) terms on different types of searchers. *Information Processing & Management* 53, 4 (2017), 851–870.

[30] Stephann Makri, Ann Blandford, and Anna L Cox. 2008. Investigating the information-seeking behaviour of academic lawyers: From Ellis's model to design. *Information Processing & Management* 44, 2 (2008), 613–634.

[31] 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.

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

[33] 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 *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 463–472.

[34] K Tamsin Maxwell and Burkhard Schafer. 2008. Concept and Context in Legal Information Retrieval.. In *JURIX*. 63–72.

[35] Douglas W Oard and William Webber. 2013. Information retrieval for e-discovery. *Information Retrieval* 7, 2-3 (2013), 99–237.

[36] Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2019. A Summary of the COLIEE 2019 Competition. In *JSAI International Symposium on Artificial Intelligence*. Springer, 34–49.

[37] D. E. Rose and R. K. Belew. 1989. Legal Information Retrieval a Hybrid Approach. In *Proceedings of the 2nd International Conference on Artificial Intelligence and Law* (Vancouver, British Columbia, Canada) *(ICAIL '89)*. Association for Computing Machinery, New York, NY, USA, 138–146. https://doi.org/10.1145/74014.74033

[38] Tony Russell-Rose, Jon Chamberlain, and Leif Azzopardi. 2018. Information retrieval in the workplace: A comparison of professional search practices. *Information Processing & Management* 54, 6 (2018), 1042–1057.

[39] Manavalan Saravanan, Balaraman Ravindran, and Shivani Raman. 2009. Improving legal information retrieval using an ontological framework. *Artificial Intelligence and Law* 17, 2 (2009), 101–124.

[40] C James Scheirer, William S Ray, and Nathan Hare. 1976. The analysis of ranked data derived from completely randomized factorial designs. *Biometrics* (1976), 429–434.

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

[42] Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. BERT-PLI: Modeling Paragraph-Level Interactions for Legal Case Retrieval. In *IJCAI-20*. 3501–3507.

[43] John I Tait. 2014. An introduction to professional search. In *Professional search in the modern world*. Springer, 1–5.

[44] Vu Tran, Minh Le Nguyen, and Ken Satoh. 2019. Building Legal Case Retrieval Systems with Lexical Matching and Summarization using A Pre-Trained Phrase Scoring Model. In *ICAIL'19*. 275–282.

[45] Marc Van Opijnen and Cristiana Santos. 2017. On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law* 25, 1 (2017), 65–87.

[46] Suzan Verberne, Jiyin He, Udo Kruschwitz, Gineke Wiggers, Birger Larsen, Tony Russell-Rose, and Arjen P de Vries. 2019. First international workshop on professional search. In *ACM SIGIR Forum*, Vol. 52. ACM New York, NY, USA, 153–162.

[47] Alice J Vollaro and Donald T Hawkins. 1986. End-User Searching in a Large Library Network: A Case Study of Patent Attorneys. *Online* 10, 4 (1986), 67–72.

[48] Ryen W White, Susan T Dumais, and Jaime Teevan. 2009. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the second ACM international conference on web search and data mining*. 132–141.

[49] Wikipedia contributors. 2021. Law of the People's Republic of China — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Law_of_the_People%27s_Republic_of_China&oldid=999943814. [Online; accessed 10-February-2021].

[50] Weijing Yuan. 1997. End-user searching behavior in information retrieval: A longitudinal study. *Journal of the American society for information science* 48, 3

(1997), 218–234.

[51] Fu Yulin. 2003. Separation of Complicated and Simple Civil Cases and Procedural Guarantees in Civil Proceedings [J]. *Cass Journal of Law* 1 (2003).

[52] Fan Zhang, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Min Zhang, and Shaoping Ma. 2020. Cascade or Recency: Constructing Better Evaluation Metrics for Session Search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 389–398.

[53] Xiangmin Zhang, Michael Cole, and Nicholas Belkin. 2011. Predicting users' domain knowledge from search behaviors. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 1225–1226.

[54] Yukun Zheng, Jiaxin Mao, Yiqun Liu, Cheng Luo, Min Zhang, and Shaoping Ma. 2019. Constructing click model for mobile search with viewport time. *ACM Transactions on Information Systems (TOIS)* 37, 4 (2019), 1–34.