Check for updates

An Intent Taxonomy of Legal Case Retrieval

YUNQIU SHAO, Department of Computer Science and Technology, Quan Cheng Laboratory, Institute for Internet Judiciary, Tsinghua University, China
HAITAO LI, Department of Computer Science and Technology, Quan Cheng Laboratory, Institute for Internet Judiciary, Tsinghua University, China
YUEYUE WU*, Department of Computer Science and Technology, Quan Cheng Laboratory, Institute for Internet Judiciary, Tsinghua University, China
YIQUN LIU, Department of Computer Science and Technology, Quan Cheng Laboratory, Institute for Internet Judiciary, Tsinghua University, China
QINGYAO AI, Department of Computer Science and Technology, Quan Cheng Laboratory, Institute for Internet Judiciary, Tsinghua University, China
JIAXIN MAO, Gaoling School of Artificial Intelligence, Renmin University of China, China
YIXIAO MA, Department of Computer Science and Technology, Quan Cheng Laboratory, Institute for Internet Judiciary, Tsinghua University, China
SHAOPING MA, Department of Computer Science and Technology, Quan Cheng Laboratory, Institute for Internet Judiciary, Tsinghua University, China

Legal case retrieval is a special Information Retrieval (IR) task focusing on legal case documents. Depending on the downstream tasks of the retrieved case documents, users' information needs in legal case retrieval could be significantly different from those in Web search and traditional ad-hoc retrieval tasks. While there are several studies that retrieve legal cases based on text similarity, the underlying search intents of legal retrieval users, as shown in this paper, are more complicated than that yet mostly unexplored. To this end, we present a novel hierarchical intent taxonomy of legal case retrieval. It consists of five intent types categorized by three criteria, i.e., search for *Particular Case(s)*, *Characterization*, *Penalty*, *Procedure*, and *Interest*. The taxonomy was constructed transparently and evaluated extensively through interviews, editorial user studies, and query log analysis. Through a laboratory user study, we reveal significant differences in user behavior and satisfaction under different

*Corresponding author

Authors' addresses: Yunqiu Shao, shaoyq18@mails.tsinghua.edu.cn, Department of Computer Science and Technology, Quan Cheng Laboratory, Institute for Internet Judiciary, Tsinghua University, Beijing, China; Haitao Li, liht22@mails.tsinghua.edu.cn, Department of Computer Science and Technology, Quan Cheng Laboratory, Institute for Internet Judiciary, Tsinghua University, Beijing, China; Yueyue Wu, wuyueyue1600@gmail.com, Department of Computer Science and Technology, Quan Cheng Laboratory, Institute for Internet Judiciary, Tsinghua University, Beijing, China; Yiqun Liu, yiqunliu@tsinghua.edu.cn, Department of Computer Science and Technology, Quan Cheng Laboratory, Institute for Internet Judiciary, Tsinghua University, Beijing, China; Qingyao Ai, aiqy@tsinghua.edu.cn, Department of Computer Science and Technology, Quan Cheng Laboratory, Institute for Internet Judiciary, Tsinghua University, Beijing, China; Jiaxin Mao, maojiaxin@gmail.com, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China; Yixiao Ma, ma-yx16@tsinghua.org.cn, Department of Computer Science and Technology, Quan Cheng Laboratory, Institute for Internet Judiciary, Tsinghua University, Beijing, China; Shaoping Ma, msp@tsinghua.edu.cn, Department of Computer Science and Technology, Quan Cheng Laboratory, Institute for Internet Judiciary, Tsinghua University, Beijing, China.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
1046-8188/2023/9-ART $15.00
https://doi.org/10.1145/3626093

Just Accepted

search intents in legal case retrieval. Furthermore, we apply the proposed taxonomy to various downstream legal retrieval tasks, e.g., result ranking and satisfaction prediction, and demonstrate its effectiveness. Our work provides important insights into the understanding of user intents in legal case retrieval and potentially leads to better retrieval techniques in the legal domain, such as intent-aware ranking strategies and evaluation methodologies.

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

Additional Key Words and Phrases: legal case retrieval, search intent, taxonomy, user behavior, user satisfaction

1 INTRODUCTION

Legal case retrieval plays a crucial role in modern legal systems [52]. In countries that follow case law system, judges rely on previous judgments of relevant cases to reach a final decision [49]. In countries with statutory law system, extensive examination of pertinent cases is conducted when presenting a case to the court to prevent erroneous judgments [16]. With the rapid growth of digitalized case documents, legal case retrieval has attracted increasing attention in both IR and legal communities. Existing research efforts [4, 34, 39, 64] on legal case retrieval mostly focus on estimating and measuring case similarity. For instance, the CAIL2019-SCM [64] task focuses on comparing the similarity between cases in each case triplet. The COLIEE [39] and AILA [4] benchmarks are designed to evaluate retrieval systems' ability in identifying supporting cases regarding a query case. However, as shown in this paper, the application scenario of legal case retrieval is broader and more complicated than similar case matching. Without knowing the actual needs of legal search users, it is difficult to develop a legal case retrieval system that is effective and reliable.

In fact, how to understand and model search intents has been a fundamental research question for IR research [6, 7, 23, 40, 55, 65]. For example, the popular taxonomy of Web search intents proposed by Broder [6] (i.e., navigational, informational, and transactional) has been widely used in the interpretation of user behavior and the design of retrieval models. It has profound implications for subsequent researches in both algorithm and evaluation design. Under different search intents, the user's expected results, search behavior, and satisfaction perception can be different significantly [5, 8, 18, 63]. Therefore, methodologies in search systems, including relevance criteria, ranking strategies, and evaluation metrics, must be adapted to different search intents accordingly [5, 8, 20, 36, 60].

The legal case retrieval scenario differs from general Web search remarkably. Specifically, the users of legal case retrieval are mainly legal practitioners with professional knowledge. The retrieved results are primarily authoritative case documents containing rich legal knowledge rather than web content with different quality levels. Instead of Web search engines, professional legal search tools (e.g., *WestLaw*, *LexisNexis*) are preferred [2, 3]. Recent research [47] has also pointed out that users' search behavior in legal case retrieval differs significantly from that in Web search. Therefore, domain-specific characteristics should also be considered regarding the search intents in legal case retrieval. However, to our best knowledge, there still lacks a well-defined taxonomy of search intents in legal case retrieval. Existing taxonomies in legal information systems are mainly designed based on legal issues and topics, such as the "Key Number System" [53], while the underlying user intents are not sufficiently studied.

Toward a legal case retrieval system that can better satisfy diverse user information needs, this paper takes an in-depth investigation into user intents. Specifically, our research questions are:

- **RQ1:** *What are the types of user intent in legal case retrieval?*
- **RQ2:** *How does user search behavior change with search intents in legal case retrieval?*
- **RQ3:** *What are the differences in perception and measurement of user satisfaction under different search intents?*
- **RQ4:** *How can the taxonomy benefit downstream tasks in legal case retrieval?*

Just Accepted

Regarding **RQ1**, we proposed a hierarchical intent taxonomy of legal case retrieval, which integrates IR and legal classification theory. To construct the taxonomy, we inspected the user surveys collected from legal practitioners and real-life queries issued to the commercial legal case retrieval engine. The taxonomy was further verified through interviews and editorial user studies. We also present the distributions of intents in legal case retrieval. To the best of our knowledge, it is the first intent taxonomy designed for legal case retrieval.

To address the above RQs, we conducted a laboratory user study with participants majoring in law. Rich behavioral data were logged to inspect the search process under different search intents. Besides, we collected explicit user feedback, such as user satisfaction and clicked reasons, to understand how users' perceptions of satisfaction change with different search intents. We also shed light on evaluating legal case retrieval across different search intents based on online metrics. Furthermore, we applied the intent categories to different downstream IR tasks, such as satisfaction prediction and result ranking. Our results reveal the significant impacts of the intent taxonomy on legal case retrieval.

To summarize, our key contributions are as follows:

- We propose a novel intent taxonomy of legal case retrieval. The taxonomy has five intent categories, i.e., search for *Particular Case(s)*, *Characterization*, *Penalty*, *Procedure*, and *Interest*. To our best knowledge, it is the first taxonomy that categorizes users' search intents in legal case retrieval.
- The taxonomy was constructed and evaluated extensively using multiple resources, such as interviews, editorial user studies, log analysis, etc. We provide the formal procedure of taxonomy creation. Moreover, we reveal the distributions of different search intents in the realistic search scenario of legal case retrieval.
- We collected a behavioral dataset with user satisfaction feedback under the proposed intent taxonomy via a controlled laboratory user study. We show significant differences in multiple search behavior patterns with different search intents. The dataset[1] has been open to the public.
- Regarding user satisfaction, we illustrate significant differences in users' perceptions of satisfaction and the different influential factors under different search intents.
- We applied the intent taxonomy to common downstream tasks, including satisfaction prediction and result ranking. Experimental results demonstrate its benefits and effectiveness.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 provides an overview of the proposed intent taxonomy in legal case retrieval. Section 4 describes the construction procedure of the taxonomy. Section 5 focuses on answering **RQ2** and **RQ3**, which introduces the user study settings and findings in user behavior and satisfaction. Section 6 presents the taxonomy's applications to satisfaction prediction and ranking tasks regarding **RQ4**. Section 7 discusses the main findings in this paper and significant implications. Finally, Section 8 discusses the conclusions, potential limitations, and future work directions.

2 RELATED WORK

2.1 Search Intent Taxonomy

It is a fundamental task for IR systems to understand users' search intents to satisfy diverse types of information needs. In Web search, Broder [6] proposed a widely adopted taxonomy using a user survey and analysis of query logs in AltaVista. According to the "need behind query", the taxonomy classified web queries into three categories, navigational, informational, and transactional. Based on this taxonomy, Rose and Levinson [40] proposed a more precise classification framework from the perspective of understanding why users are searching. Recently, Cambazoglu et al. [7] built a new multi-faceted intent taxonomy for questions asked in Web search engines based on 1,000 real-life issued questions, which was more fine-grained but less ambiguous for human assessors. Bolotova et al. [5] presented a comprehensive taxonomy of the current non-factoid question-answering task and

[1] https://github.com/THUIR/Search-behavior-dataset

Just Accepted

they also pointed out that the challenging categories were poorly represented in the existing datasets. Besides web search, search intents have also been investigated in some specific search scenarios, such as multi-media search [23], image search [33, 65], product search [51, 55], medical search [58], etc. Given different search intents, some research indicated that user search behavior would also be different [18, 55, 63, 65]. Meanwhile, with a good understanding of underlying search intents and related effects, search engines could be further improved in various aspects, such as diversity search [8], personalized search [60], query suggestion [20], result ranking [15], satisfaction prediction [36], and system evaluation [5, 68].

Meanwhile, taxonomies in the legal field are almost centered on objective legal knowledge, such as classifying law systems [50], rules of law [48], and legal issues [53]. For example, the well-known "Key Number System" in Westlaw [53] is a kind of taxonomy that organize cases by legal issues and topics. However, user search intents were not included in these taxonomies. To the best of knowledge, there is no systematic modeling of search intents in legal case retrieval.

2.2 Legal Case Retrieval

Legal case retrieval is a specialized IR task that aims to search for relevant legal cases given the matter at hand. Compared to web search, legal case retrieval has unique challenges.On the one hand, legal documents are often much longer and use domain-specific terminology than web document. On the other hand, the definition of relevance for legal case retrieval goes beyond simple semantic similarity. It is a crucial task in legal practice and has drawn active research efforts in both legal and IR communities. In the earlier decades, extensive expert efforts were invested in organizing legal knowledge and developing the professional legal information system. For example, Moens [37] identified some form of concept based retrieval, containing three models: Boolean retrieval, vector space retrieval, and probabilistic retrieval. Klein et al. [22] outlined ontological-based approaches for retrieving similar cases to a seed case. In recent years, with the rapid increase of digitalized case documents and the development of NLP techniques, research efforts have been put into developing automatic retrieval models that can identify relevant cases given a query case. Several benchmarks have been constructed for this task, such as COLIEE [39], AILA [4], CAIL2019-SCM [64], and LeCaRD [34], where the core concern is to measure the semantic relationship (e.g., similarity) between cases. For instance, in CAIL2019-SCM [64], the task is to detect which two cases are more similar in each triplet. Based on these benchmarks, a variety of case retrieval models [26, 29, 35, 45, 59, 67] have been proposed, such as measuring case relevance via automatic summarization [41, 57], paragraph semantic modeling [45, 59], rationale matching [67], and so on. LOCKE et al. [32] summarize the methods of case law retrieval in the past 30 years and point out that the future of case law retrieval is based on natural language. For example, Savelka et al. utilize pre-trained language models to discover explanatory sentences for legal cases [42–44]. Li et al. utilize the structure of legal cases to design new pre-training objectives, which yielded state-of-the-art results on legal case retrieval [26, 28]. Beyond developing semantic case-matching models, some recent works pay attention to user-system interactions in legal case retrieval. In particular, Shao et al. [47] conducted a user study to investigate user behavior in legal case retrieval and illustrated significant differences from the web search scenario, e.g., the exploratory property. Given the task complexity in legal case search, Liu et al. [30] attempted to apply the conversational search paradigm, which might facilitate the users expressing their information needs better, according to their user study. It is noteworthy that the tasks in these studies were still designed based on the classification of objective legal knowledge, such as the rules of law [45] or the legal issues [30]. However, as far as we know, users' underlying search intents in legal case retrieval have not been investigated systematically nor considered in existing studies, such as retrieval model development or user study design.

Just Accepted

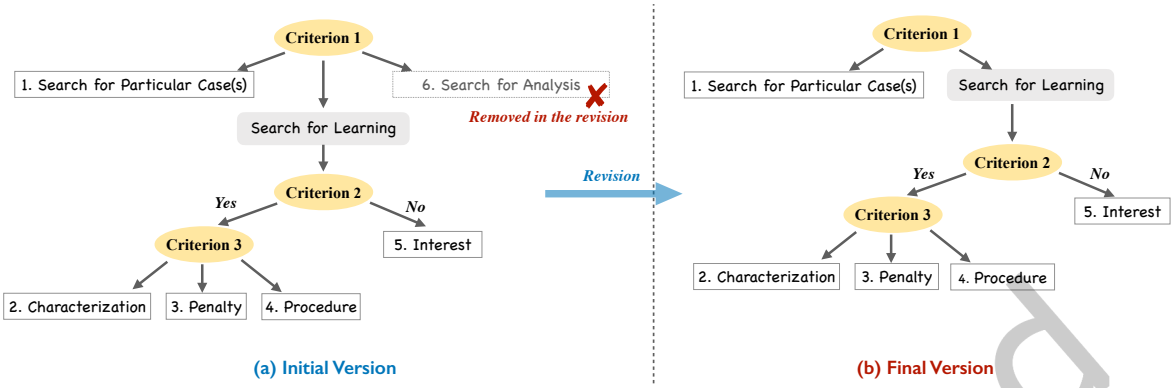


Fig. 1. Illustrations of the proposed taxonomy structures. The (a) is the initial version of taxonomy and the (b) is the final version.

3 TAXONOMY OVERVIEW

In this paper, we focus on the hierarchical intent taxonomy for legal case retrieval. We attempt to construct and validate the intent taxonomy in the Chinese legal system. Specifically, the judicial process in China consists of four steps: prosecution, acceptance, preparation for trial, and trial. After the trial, the court will write a legal case document, consisting of basic information such as facts, reasoning, and the judgment. These legal case documents can be of great help to users. Legal practitioners can retrieve similar cases to help make decisions, and ordinary people can learn more about the law from legal cases. Our taxonomy is oriented to real retrieval scenarios. In real scenarios, users retrieve similar cases in order to solve the matter at hand. This means that the form of the question can be varied in a particular intent. In short, the corpus of the legal case retrieval is a large number of legal documents, while the query may be a case, a sentence or some keywords.

There is no doubt that there is search intent when users enter a query in the legal case retrieval, i.e., a certain aspect of the case is expected to be retrieved. This intent may be explicit or implicit. A good case retrieval system should facilitate the users to express different intents and return appropriate cases. Therefore, to better guide the design of the case retrieval system, we propose a novel hierarchical intent taxonomy of legal case retrieval, which integrates IR and legal classification theory. This section gives an overview of the final intent taxonomy, as shown in Table 1 and Figure 1. Figure 1 (the right one) illustrates the general framework of the proposed intent taxonomy. Specifically, the following three criteria are utilized to categorize user intents successively.

- **Criterion 1** What is the purpose of legal case retrieval?
- **Criterion 2** Is the search driven by a clear objective or not?
- **Criterion 3** What kind of legal problem does the objective belong to?

The first criterion (**Criterion 1**) asks what users are searching for. According to this criterion, intents can be generally classified into two groups, search for *Particular Case(s)* (PC) and search for *Learning* (Le) from the cases. The *PC* intent can be compared to a combination of navigational and transactional needs in Web search [6] or a known-item search in product search [55]. Meanwhile, the *Learning* category is somewhat similar to an informational need. Furthermore, the category *Learning* involves multiple situations in legal case retrieval, and thus we apply the second criterion (**Criterion 2**), concerning whether there is a clear objective to learn [40, 65]. If not, we consider that the search session is to satisfy some individual interest, such as curiosity or gossip triggered by social news. Otherwise, we apply the third criterion (**Criterion 3**) to categorize the specific objective based on

Just Accepted

Table 1. The proposed intent categories in legal case retrieval and examples. The *Analysis* category was removed in the taxonomy revision process. It's worth noting that the Example is not a query that the user enters into the search engine, but rather a problem that they are currently attempting to address.

Category	Description	Example
1. Particular Case(s) (PC)	Search for some particular case document(s), e.g., the judgment documents of a specific case, the parties' previous convictions or lawsuits.	What are the lawsuits that Company A has involved?
2. Characterization (Ch)	Search for learning about conviction or law application under the substantive law regarding the current issue. With this intention, users focus on the characteristics of different aspects of the underlying facts, such as different case types, causes, regions, statutes etc.	Whether trapping loans is constituted fraud? Whether the claim is based on product liability or consumer fraud liability?
3. Penalty (Pe)	Search for learning about sentencing or penalty range regarding the current issue. Under this intent, users focus on the violations and corresponding penalties involved in the case, such as criminal penalties (imprisonment, probation), civil damages (economic damages, moral injury), administrative penalties (fines, revocation of business licenses), etc.	What is the punishment for embezzling $100,000 in XX? Whether the request for the return of $7,700 and punitive damages can be supported
4. Procedure (Pr)	Search for learning about procedure issues related to the procedural law, i.e., litigation procedures, appeal procedures, evidence collection procedures, and enforcement procedures.	What procedure should be followed if an undergoing civil case involves a criminal offense? Whether acts committed in 2017 are time-barred?
5. Interest (In)	Have no specific legal issue to solve but search for learning some related information to satisfy the individual interest.	Johnny Depp v. Amber Heard; What are the recent cases that apply the XX rule?
6. Analysis	*Search for writing some analytical reports, e.g.,similar case search report, statistical survey on a specific charge.*	*Writing a similar case search report regarding the XX case; An empirical study of corruption based on over 200 judgments.*

Just Accepted

the general classification of law [61], i.e., substantive law[2] or procedural law[3]. Specifically, substantive law refers to the set of laws that governs how members of a society are to behave. In contrast, procedural law (also referred to as adjective law) comprises the rules of procedures for making, administering, and enforcing substantive law. Note that this classification exists in different law systems, in other words, is generally applicable across law systems. Based on **Criterion 3**, we group search intents into three categories, *Characterization* (Ch), *Penalty* (Pe), and *Procedure* (Pr). The *Procedure* intent is about issues under the procedural law. The *Characterization* and *Penalty* are classified based on two primary types of issues under the substantive law, such as crimes and punishments. Table 1 provides the descriptions and examples of each intent category. It is worth noting that the example in

[2]https://en.wikipedia.org/wiki/Substantive_law
[3]https://en.wikipedia.org/wiki/Procedural_law

table 1 is not a query entered by the user but rather the intent. Users can construct various forms of queries to realize their search intent according to their own customs. For example, when a lawyer is dealing with a legal case, he would like to know the possible penalties for the defendant's behavior. Then he can input several keywords or this case into the search engine.

Given different search intents, the results that users want to retrieve could have different properties. According to **Criterion 1**, users would have strong needs for both precision and recall under the *PC* intent. The potential relevant cases might be definite and of a limited size, correspondingly. On the contrary, the relevant cases under *Learning* could be broader. Furthermore, according to **Criterion 2**, one could expect that search under the *Interest* intent would have relatively lower requirements on precision and recall compared to the others. Among the three types categorized by **Criterion 3**, *Characterization* and *Penalty* are based on the substantive law, while the *Procedure* focuses on the issues under the procedural law. Therefore, the relevance criteria under the *Procedure* might differ from those used in *Characterization* and *Penalty*. Meanwhile, comparing *Penalty* with *Characterization*, the information need under *Penalty* would be more specific and precise, and thus precision would be more emphasized than under *Characterization*.

In summary, we construct an intent taxonomy based on three criteria. The taxonomy consists of five intent categories, i.e., Search for *Particular Cases* (PC), *Characterization* (Ch), *Penalty* (Pe), and *Procedure* (Pr), and *Interest* (In), and the detailed construction process is described in the next section. It is worth noting that this legal classification theory mainly takes the Chinese legal system into consideration. We primarily verify the rationality of the taxonomy under the Chinese legal system. We believe that it can contribute to the legal community and inspire the development of taxonomies for different legal systems.

4 CONSTRUCTION PROCEDURE

The taxonomy was constructed and evaluated extensively in an iterative way. In this section, we describe the creation procedure as illustrated in Figure 2. Generally, the procedure can be divided into two stages, *I. establishment*, and *II. revisit and verification*. In the establishment stage, we propose an initial version of the taxonomy. Then, we move on to verifying the taxonomy through different methods and revising it accordingly before settling down the final version.

4.1 Establishment

To construct the initial intent taxonomy, we exploited three resources, including the literature, user survey, and sampled query logs in the *establishment* stage. We studied the literature on taxonomy in information retrieval and classification theory in the legal domain. In addition, we inspected users' real-life information needs in legal case retrieval through a user survey and query logs.

User Survey. We designed an online survey to collect users' recent legal case retrieval experiences. Besides basic demographic questions, each participant was asked to answer the following three open questions according to her last time of legal case retrieval:

(1) What is the context that triggers this search? (e.g., working-on case, undergoing research topic, others' consultation, social news or hotpots, etc.)
(2) What is the detailed task of this search? (Specific queries and query intents, e.g.,client's litigation situation, equity betting cases, etc.)
(3) What search engine(s) did you utilize in this search?

The survey was spread via social media platforms, such as WeChat, etc. Since the target users of legal case retrieval are mainly legal practitioners, we only collected responses from participants engaged in law-related occupations and paid each participant about $1 for her serious response. We received responses from 116 participants and kept 110 after filtering the answers that were too vague or unrelated to legal case retrieval.

Just Accepted

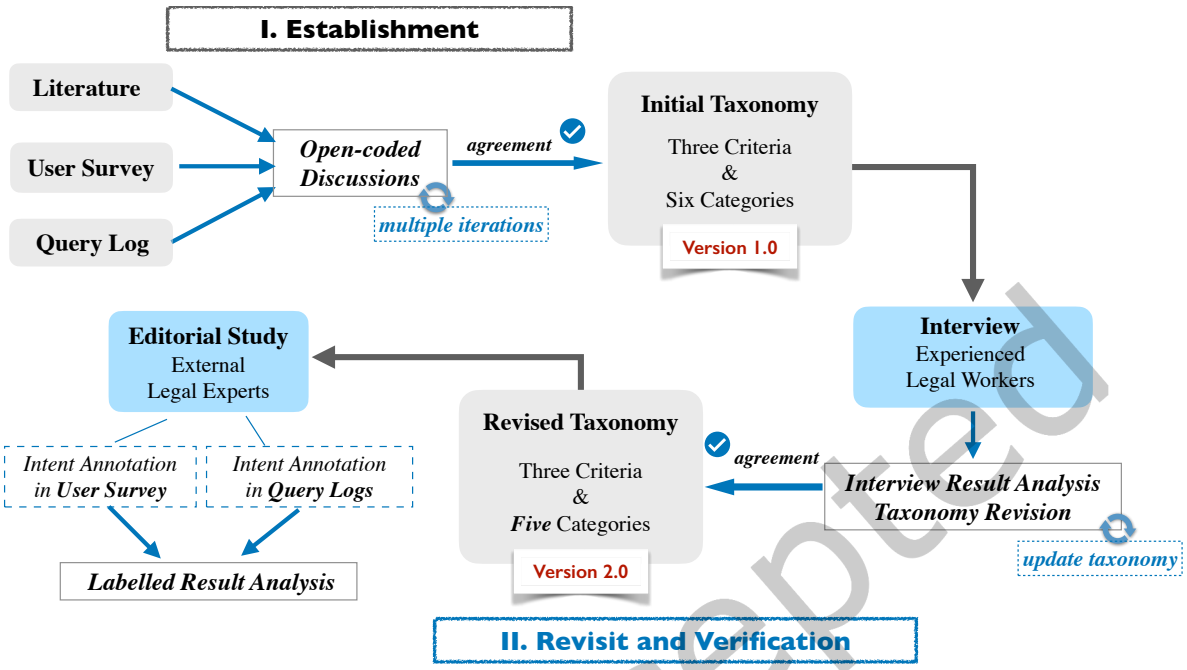


Fig. 2. The procedure of taxonomy creation, which consists of two stages (i.e., "I. establishment" and (2) "II. revisit and verification') in general. The second stage includes two parts, (1) semi-structured interviews and (2) an editorial user study.

Table 2. Occupational distribution of user survey participants. Court, Procuratorate, and Corporate represent the staff of each of them.

Occupation	Court	Procuratorate	Lawyers	Corporate	Legal Researcher	Other
Number	10	17	31	18	14	20
Ratio	9.09%	15.45%	28.18%	16.36%	12.73%	18.18%

The participants were from various legal-related occupations, including lawyers, staff of corporate legal affairs, prosecutors, judges and court staff, and legal researchers. Table 2 shows the occupational distribution of the participants. The user survey helped us gain a deeper understanding of the diverse search intents and tasks performed by legal professionals in their daily work.

Query Logs. We sampled 600 search sessions from a commercial legal case search engine[4]. They were sampled from 7-day query logs in August 2021, involving 516 users. Similar to previous studies [65], we used 30 minutes as the window for splitting sessions. Furthermore, we excluded the sessions with less than one query term, which might be too vague to identify the search intents [47]. Then, we randomly sampled 100 of them while establishing the initial taxonomy. Following [65], we assume each session to involve one topic. The query log analysis provided valuable data on real-world search behaviors, allowing us to identify prevalent search patterns and refine our understanding of user intents.

[4]https://ydzk.chineselaw.com/case

Just Accepted

Initial Taxonomy. The authors closely reviewed the survey responses and query logs to induce potential criteria for classifying user intents. Following previous work [40, 65], we took several iterations of the open-coded discussion before reaching an agreement. The initial version of the taxonomy is as shown in Figure 1 (the left one), which consists of six intents categorized by three criteria. The three criteria are the same as described in Section 3. Specifically, the first criterion leads to three categories, i.e., Search for *Particular Case(s)*, *Learning*, and *Analysis*, in this version. The *Analysis* category denotes searching for writing some analytical reports, such as similar case search reports, which are sometimes required in the judicial process. At this point, we considered it as an independent search intent since it is a specialized task in judicial practice. The following section (Section 4.2) will explain how we revise and verify the intent taxonomy.

4.2 Revisit and Verification

As illustrated in Figure 2, we conducted semi-structured interviews, collecting exhaustive feedback from experienced legal practitioners and making a qualitative analysis. Revisions were made accordingly. Following that, we conducted an editorial user study based on user survey responses and query logs, making a quantitative inspect.

4.2.1 Interview. We conducted semi-structured interviews with four experienced legal workers separately, including one lawyer, one prosecutor, and two judges. They all work in Beijing. Three of them are men and one is a woman. Although the face-to-face interview only involves a small sample, it could allow a more in-depth questioning and discussion, broadening and deepening the understanding on the research problem [56]. The interviewees in our study were all well-experienced in legal practice and came from representative legal occupations. Each interview took about 30 minutes. Interviewees were compensated about 100 dollars for their participation. The audio was recorded for later analysis.

To begin with each interview, the interviewer introduced the proposed taxonomy in detail, including the three criteria, the hierarchical structure, and six intent categories, as shown in Figure 1 and Table 1. Each interview was centered on two open questions plus a short series of follow-up questions.

The first question asked about the intent taxonomy's coverage and rationality. Here is an example[5]. Firstly, we asked, *What do you think of the coverage of this taxonomy? Can it cover all your information needs in legal search daily? If not, is there anything else that needs to be added?* Then, we followed with questioning about the concrete categories. For example, *How about the XXX category? What do you think about the definition and characteristics of this category?* The follow-up question is a good trigger for open discussions. We collected rich comments and views on these intent categories from the perspectives of diverse legal occupations, which further helped us revise the taxonomy.

The second question asked about the importance of different categories in the interviewee's daily search. For example, *Among these intent categories, what do you think are more critical or occur more often in your daily search? And why?* We designed this question to obtain explicit feedback on the importance of different intents in the practice of legal case retrieval. Unlike user surveys or search logs, we could receive much more fine-grained explanations regarding this aspect despite the small data samples.

Results. After completing all the interviews, we analyzed the records. The main results are summarized as follows.

(1) Regarding the first question, all the interviewees stated that the proposed taxonomy has good coverage of daily needs in legal case retrieval. No more new categories were proposed.
(2) Regarding the comments on each intent category (i.e., the follow-up question), the *Analysis* category attracted plenty of discussions. The lawyer and the judges, who usually dealt with such analytical reports (e.g., similar case search reports), indicated that this type (*Analysis*) could be covered by the other categories

[5]Interviews were all in Chinese. We show the translation in this paper. The exact wording varied for each interview.

Just Accepted

mentioned earlier. Although it highly depends on the individual case, the underlying information need is still to learn about a specific legal problem (e.g., *Characterization* or *Penalty* most of the time). The prosecutor interviewee indicated that he seldom had this type of intent. The potential situation he came up with is that when dealing with a difficult legal issue (e.g., the *Procedure*), he might also sum it up to an analytical report afterward, such as personal learning material.

(3) Other comments on the concrete categories are centered on the categories under **Criterion3**. To be specific, the *Characteriztion* category should also include the situations of innocence and those of non-prosecution (*from the prosecutor*). Under the *Procedure* intent, they usually search for the legal requirement related to jurisdiction or avoidance (*from the lawyer*). Regarding the *Penalty* intent, all the interviewees mentioned that it has attracted increasing attention in recent judicial practice, but meanwhile it is much harder to be satisfied in the current legal case search systems. Precision was especially emphasized under this intent. Last but not least, they all suggested that these three intents expect more diversified results than the *Particular Case(s)* intent and meanwhile require higher precision and recall than the *Interest* intent.

(4) Regarding the second question, all the interviewees suggested that the *Characterization* and *Penalty* are the most important and common in their daily search. Especially, the *Penalty* was emphasized again by the prosecutor and the lawyer separately. Meanwhile, the prosecutor and the lawyer also mentioned that the *Procedure* is highly significant. Although the *Procedure* intent is less common than the above two categories in legal case search, it will be pretty valuable and, meanwhile, difficult if there is a need for case retrieval surrounding the procedure requirement.

Revisit and Revision. Based on the interview responses, we had further iterative discussions on the taxonomy and finally reached an agreement on the revision as illustrated in Figure 1. To be specific, we removed the *Analysis* category since it could be covered by the left intent categories. It would be better to view it as a context that triggers a legal case search rather than an independent intent search category. Furthermore, the first two authors re-coded the user survey and the sampled search logs that were used to establish the taxonomy in Section 4.1. As a result, the revised taxonomy still had good coverage. To summarize, we achieved a revised taxonomy composed of three criteria and five intent categories, as illustrated in Figure 1(b). Meanwhile, the in-depth discussions and exhaustive feedback help us further clarify the definitions of intent categories and also give us an qualitative view of the importance of different categories in practice.

4.2.2 Editorial User Study. To verify the revised taxonomy, we further conducted an editorial user study. In this study, we recruited three external legal experts to annotate the users' search intents in the user survey responses and query logs. The user survey responses and query logs are those described in Section 4.1. Unlike the establishing stage, we utilized all the sampled query logs (600 sessions in total) this time. The three annotators were all graduate students majoring in law and qualified in legal practice[6]. They all reported using legal case retrieval regularly and being familiar with current legal case search engines. They all signed a consent form before participating.

At the beginning of the study, we introduced the revised taxonomy in detail. We provided the annotators with the criteria, the taxonomy structure, the description, and examples of each intent category. In addition to the five intent categories, we provided another two choices for the annotators, *Others* (O) and *Multi* (M). The *O* means that the search intent does not belong to any of the proposed categories. The *M* denotes that the underlying intent seems to fall into multiple categories. For the additional two choices, we asked annotators to provide explanations for their choice. For example, the annotator needed to give what intent categories the search task might fall into if she selected *M*. The annotators were required to annotate the underlying search intent of each response in the user survey based on the answers to the three open questions and annotate the search intent of

[6]They had passed the "National Uniform Legal Profession Qualification Examination" and had at least five years of law-related experience

Just Accepted

Table 3. Distribution of search intent categories.

Intent Category	User Survey	Query Log
Particular Case(s) (*PC*)	7.27%	21.24%
Characterization (*Ch*)	50.00%	54.85%
Penalty (*Pe*)	10.91%	9.03%
Procedure (*Pr*)	6.36%	4.01%
Interest (*In*)	12.73%	0.17%
Others (*O*)	0.91%	0.67%
Multi (*M*)	11.82%	10.03%

each session according to its queries. After all the annotators confirmed a good understanding of the taxonomy, they annotated the survey responses and query logs independently.

It took about 1.5 hours and 7 hours on average for each annotator to annotate the user survey responses and query logs, respectively. Each annotator would be paid about $12 for a one-hour annotation. As for label aggregation, we utilized the majority vote. In particular, if every annotator made different annotations for a sample, we tagged it as *Multi* (M).

Results. The Fleiss's Kappa [12] κ among three annotators is 0.62 in terms of the user survey annotation, reaching a substantial agreement ((0.61, 0.80)). As for the query log annotation, the κ among three annotators is 0.58, reaching a moderate agreement ((0.41-0.60)). Compared with the survey where users described their search scenario explicitly, the query logs were vaguer for intent labeling and thus explained the slight drop in κ [55, 65]. Given the relatively high number of categories, the inner-annotator consistency is acceptable [5, 65] for both datasets, suggesting that the taxonomy can be easily understood and distinguished.

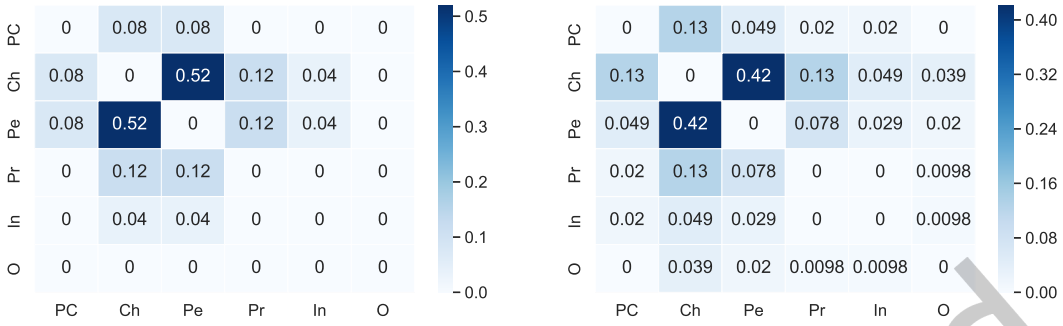
Table 3 shows the proportion of each intent category. As a result, less than 1% search tasks were annotated as *Others* in both user survey and query log datasets, indicating that the proposed taxonomy has good coverage of users' intents in legal case retrieval.

Regarding the five categories in the taxonomy, the general distributions are similar in both datasets, especially for the three intents classified by **Criterion 3**. In particular, the *Characterization* intent accounts for about 50%, indicating it is a fundamental and common task in legal case retrieval. Consistently, recent research and benchmarks [34, 46, 47] designed tasks mainly based on this category. Meanwhile, the proportions of the *Penalty* and the *Procedure* are lower but still non-trivial compared with that of the *Characterization*, which also aligned with the feedback collected in the interviews. These two are also primary tasks in the legal decision process. Besides, as the interviewees (in Section 4.2.1) also pointed out, the need for *Penalty* has been growing and is increasingly important in recent years.

Meanwhile, the distributions of *Particular Case(s)* and *Interest* vary in the two datasets. A higher proportion of *PC* intent is observed in search logs while few *Interest* intents are in the logs. We think that users' explicit responses in the survey can better reflect their real information needs, while the query logs are implicit indicators. Besides, the search engine itself might cause some bias in user preference. For example, if the search engine is not good at satisfying *Interest* needs, users may not like to use it for this intent, and vice versa. According to the survey, we also noted that some users would use Web search engines rather than legal databases under the *Interest* intent.

Mixture Analysis. Nearly 10% of search tasks are tagged as *Multi* in both datasets. Note that *Multi* in Table 1 consists of two parts, i.e., more than two annotators labeled as *Multi* (7.27% and 4.85% in the survey and logs,

Just Accepted



(a) User Survey (b) Query Logs

Fig. 3. Confusion matrix of the "Multi" in the user survey (a) and query logs (b). Each matrix is a symmetric one and the number in the grid denotes the normalized frequency of co-occurrence.

respectively) or three annotators gave completely different labels (4.55% and 5.18% in the survey and logs, respectively). To deeply analyze it, we visualize the co-occurrence of different intents, as shown in Figure 3. For each sample belonging to the first part, we manually processed the annotators' explanations, from which we extracted all potential intents. We considered all annotated categories as possible intents for each sample belonging to the second part. Then, we count each pair in the possible intent set as one co-occurrence. For example, the intent set, "Ch+Pr+Pe", contributes once occurrence for the "Ch-Pe", "Ch-Pr", and "Pe-Pr" pairs, respectively. Numbers in Figure 3 are normalized by the number of pairs.

As shown in Figure 3, the pair of *Characterization* and *Penalty* is the one that co-occurs most frequently in both user survey and query log data, accounting for around 50%, which suggests that users might search for both needs simultaneously. Meanwhile, we observe that the *Procedure* usually co-occurs with the above two intents, which also aligns with the hierarchical structure of the intent taxonomy. Generally speaking, the query logs, where user intents could only be inferred implicitly, involve more types of co-occurrence of potential intents compared to the user survey. The results here suggest retrieval methods that explicitly recognize multi-intent queries are needed.

4.2.3 Summary. Based on the *revisit and verification* stage composed of the interviews and editorial user study, we finalized the intent taxonomy, consisting of five categories, i.e., Search for *Particular Case(s)*, *Characterization*, *Penalty*, *Procedure*, and *Interest*. We also provide quantitative insights into the search intent distributions in legal case retrieval, according to the intent annotations of user survey responses and query logs.

5 SEARCH BEHAVIOR AND SATISFACTION

To understand user search behavior and satisfaction under different search intents, we conducted a laboratory user study using the proposed taxonomy. In this section, we described the behavioral data collection process and addressed **RQ2** and **RQ3** with the collected data.

5.1 User Study

5.1.1 Tasks and Participants. We designed three search tasks for each intent category. In each task, we provided a query case description as the background and a question specific to the intent category. Table 4 shows a criminal example and a civil example. The participant needed to retrieve relevant cases regarding questiones. All the tasks

Just Accepted

Table 4. Examples of User Study. The different questions reflect the corresponding type of intent. Users need to retrieve relevant cases regarding the question.

Background	Question
Criminal case: The defendant Song was the son-in-law of the victim Li, and the two had a long-standing conflict. At 3:00 p.m. on January 20, 2019, the two clashed. Song rushed into the kitchen and casually picked up a knife and stabbed Li. After Li was stabbed 8 times, the knife suddenly broke. At the same time, Song's wife Chen just came home and went up to pull to stop Song. Seeing his wife Chen's grief, Song felt very regretful and sorry for his wife, so he held a knife to self-harm. Chen immediately called the police, Song knew that Chen was calling the police and did not stop. Because of the seriousness of the injury, the police immediately took Song to the hospital, the day began residential surveillance, by two police officers at the same time.	**PC**: The proceedings of Song. **Ch:** Is the defendant's act an intentional killing or intentional injury? **Pe:** Is the defendant's behavior a criminal suspension or an attempt to commit a crime? **Pr:** Can residential surveillance be treated as a prison term?
Civil case: Zhang was running a beauty salon. between 2016 and 2019, Chen spent several times at Zhang's beauty salon and paid Zhang a total of RMB 7,700. Zhang informed Chen that all of the above items were of the best quality, but in reality they were all fake. The hyaluronic acid injected by Zhang for her caused Chen to develop hard lumps on her chin and chest. Chen filed a lawsuit in January 2020.	**PC:** The proceedings of Zhang. **Ch:** Is Chen's claim based on product liability or consumer fraud liability? **Pe:** Is Chen's request for the return of $7,700 and punitive damages supportable? **Pr:** Is the act committed by Zhang in 2017 time-barred?

Just Accepted

were adopted from the real cases to simulate the realistic search scenario. Similar to the previous study [47], we anonymized the background case and removed the courts' opinions. So the query case contains only the basic facts. All the tasks were designed to be of moderate difficulty to avoid impacts of task difficulty. Following [47], moderate difficulty cases are selected by experienced law professors. The *Interest* category was not included in the user study, since we could hardly simulate the search tasks triggered by an individual interest in a laboratory setting according to our prior study. In total, we designed 12 search tasks for the left four categories (i.e., *PC*, *Ch*, *Pr*, and *Pe*) and an additional warm-up task.

We recruited 36 participants that were qualified[7] in legal practice. Specifically, eight were lawyers and the others were students in law school. They were native Chinese speakers and familiar with legal case retrieval. Considering the workload, we assigned each participant 3 main search tasks along with a warm-up training. On average, it took about 1.5 hours for each participant to complete the tasks. We carefully designed the assignment that each participant completed tasks of three different intents, and each task was completed by nine different participants. Tasks were shown in a random order to balance the order effects [25, 47]. A pivot study involving two additional participants was conducted ahead to ensure the experimental design worked well.

5.1.2 Procedure. First of all, we introduced the entire experimental procedure. After signing the consent form, the participant was directed to a warm-up task to get familiar with the experimental settings. Then, the participant moved on to the three main tasks. Each task consists of the following four steps.

Step1. The participant was provided with the query case description and the question. She was instructed to search for relevant cases that could help her answer the question. After reading the case background and the

[7]They passed the "National Uniform Legal Profession Qualification Examination"

Table 5. Descriptions of options in the user study (Step4).

Option	Description
Relevance	The relevance to the query. If the retrieved case satisfies the user's search intent, it is considered relevant otherwise irrelevant.
Diversity	The diversified content or opinions, e.g., providing different information or opinions beyond the cases that have been retrieved.
Authority	The authority of the retrieved case, e.g., the court level involved in the case.
Timeliness	The time-related factors of the retrieved case, e.g., the time that the case happened or was judged.
Region	The region-related factors of the retrieved case, e.g., the region that the case happened or was judged.
Inspiration	The inspiration of the result, e.g., providing ideas of identifying useful cases or formulating queries.
Ranking	The ranking position of the result.

question, the participant filled in a pre-search questionnaire to report her perceived task difficulty on a 5-point Likert-type scale.

Step2. The participant was directed to the experimental search engine. The participant could conduct legal case retrieval freely as she usually did, such as querying, clicking, turning pages, and so on. The participant could finish the search session once she found enough results or could not find more.

Step3. The participant was directed to a post-task questionnaire that contained two questions. The first is to ask for her perceived satisfaction regarding the entire search session on a 5-point scale. The second requires the participant to summarize the retrieved results and answer the task question. This question is to ensure the participant accomplished the search tasks seriously.

Step4. The participant was instructed to provide feedback for each query. Specifically, the issued queries, along with the questionnaires, would be shown successively. Regarding each query, the SERP and titles of clicked results (if any) were also provided for reminding. The participant was instructed to report her satisfaction on a 5-point scale (1:not at all, 5:satisfied) regarding this query and select the reasons for clicking on the results (if any). The reasons for clicking were collected in the form of a multi-choice question. The options include *relevance*, *diversity*, *authority*, *timeliness*, *region*, *inspiration*, *ranking*, and *others*. The descriptions of options are as shown in Table 5. If the participant chose the *others* option, she also needed to provide the potential factors that were not included in these options.

5.1.3 Experimental System. We developed an experimental platform using Django where the participants completed the entire study procedure. As for the experimental search engine, we redirected to a commercial legal case retrieval engine [8]. Query suggestions and advertisements were filtered. It had been confirmed in advance that the search system would not do personalization. We developed a customized chrome extension to log user behavior and examined pages.

[8]https://ydzk.chineselaw.com/case

Just Accepted

Table 6. Differences in user behavior with different search intents. PC/Le/Ch/Pe/Pr denotes for Particular Case(s), Learning, Characterization, Penalty, and Procedure, respectively. "*/**/***" indicates a significant difference at $p < 0.05/0.01/0.001$ level (after Bonferroni-Holm correction).

Group	Behavioral Measure	Criterion 1			Criterion 3			
		PC	Le	sig.	Ch	Pe	Pr	sig.
Task Events	# query per session	6.346	6.870	–	4.615	**9.000**	7.000	*
	# pages per session	11.17	13.78	–	10.24	**17.74**	13.32	**
	# search depth in pages	**1.309**	1.063	**	1.087	1.052	1.029	–
Click	# clicks per session	5.615	5.886	–	3.435	**7.593**	4.565	***
	min click rank per query	1.269	**1.963**	***	1.630	1.913	2.500	–
	avg click rank per query	2.823	**3.551**	*	3.264	3.288	**4.189**	*
	% sats click per query	0.3294	**0.6485**	***	**0.7150**	0.7090	0.5214	**
Hover	# hovers per session	50.78	46.73	–	33.61	**59.44**	46.95	*
	min hover rank per query	1.142	1.075	–	1.095	1.051	1.088	–
	avg hover rank per query	3.059	3.170	–	3.356	3.171	3.062	–
	avg hover time (seconds) per query	2.221	**2.790**	**	2.722	2.938	2.684	–
	P(click|hover) per query	0.2044	0.1817	–	**0.2099**	0.1927	0.1484	**
Dwell Time	task time (seconds) per session	379.5	**551.3**	**	438.3	**702.7**	425.8	**
	% SERP time per session	**0.6042**	0.4218	***	0.3868	0.4255	0.4637	–
	avg click dwell (seconds) per query	28.76	**53.37**	***	**66.83**	53.10	43.41	**

5.1.4 Dataset. We collected 108 valid search sessions (843 queries included) for the 12 tasks from 36 participants after the quality check. The dataset contained rich behavioral data (e.g., queries, clicks, hovers, and timestamps) and users' explicit feedback (e.g., satisfaction and click-through reasons).

5.2 Search Behavior under Different Intents

Regarding **RQ2**, we investigated users' search behavior under different intents based on the collected behavioral data.

The search tasks were designed to be of similar difficulty across different intents to avoid the potential influences of task difficulty [47]. This design is also verified by users' feedback on task difficulty in *Step1*. The average task difficulty is 2.5, and no significant differences were observed across intents ($p > 0.6$).

Search intents are considered as independent variables. We follow the hierarchical structure in Figure 1 for an in-depth investigation. Specifically, we first group sessions into the *Particular Case(s)* and *Learning* categories according to **Criterion 1**. Then, we apply **Criterion 3** to sessions in *Learning* and group them into the *Characterization*, *Penalty*, and *Procedure* categories. We investigate user behavior in legal case retrieval from multiple aspects, including task events, click, hover, and dwell time. Non-parametric statistical tests (Kruskal-Wallis test [24]) are utilized since these measures have non-normal distributions (K-S test). The p-values are calibrated through Bonferroni-Holm adjustment [17] within the behavioral group to counteract the multiple comparison problem [14]. Results are given in Table 6.

Task Events. Comparing the *Particular Case(s)* and *Learning* intents, the general numbers of issued queries and visited pages are similar. We suppose this may be due to the fact that users tend to prefer simple keyword

Just Accepted

expressions when searching for particular cases. When the satisfactory case cannot be found using simple keywords, the user will further enrich the query description. Regarding search depth in pages (the number of SERP pages a user browses per query [55]), users turn pages more often under the *Particular Case(s)* intent. On one hand, this reflects the fact that current user search habits do not facilitate the rapid identification of specific cases. On the other hand, the difference reflects the requirement for both high precision and high recall given the *Particular Case(s)* intent. Meanwhile, we observe significant differences in the number of queries and pages when comparing among *Characterization*, *Penalty*, and *Procedure*. More queries and pages are examined under the *Penalty* intent, indicating a higher search effort. The results could be interpreted by the legal characteristics of the *Penalty* category that the underlying information need is usually more specific and precise.

Click. We observe significant differences in all the query-level click-through measures between the *Particular Case(s)* and *Learning*. The difference in search purpose (**Criterion 1**) might lead to remarkably different examination patterns within a query. The differences in *min* and *avg* clicked positions indicate that users with *Learning* intent seem more patient and careful with the returned results. When using a 30-second threshold to determine a satisfactory click, a higher proportion of clicks under the *Learning* intent category meet this satisfaction criteria. Comparing among the intents categorized by **Criterion 3**, click-through behavior measures also vary significantly. The *Penalty* involves the most clicks within a session, which is consistent with the analysis of the above task event measures. Furthermore, according the *%sats click*, users seem to be least satisfied with the results under the *Procedure* intent. Different from *Characterization* and *Penalty*, the requirement under *Pr* is based on the procedural law and the corresponding relevance criteria might also differ. Without understanding the user intent, the existing retrieval systems might hardly resolve this kind of information need well.

Hover. Following previous works [10, 47], we utilize hover measures to reflect users' examination of the results shown on SERPs (e.g., snippets). They could capture more behavioral information than click-through measures but might involve more noise. Specifically, we view hover-through as a signal of a preliminary examination, which involves less examination effort than click-through. Comparing *Particular Case(s)* and *Learning*, the main difference lies in the average hover time on results. In the context of *Learning* intent, it would take users more time to examine and understand the result content. Among the three categories under *Learning*, *Penalty* involves the most hovers, which may indicate a higher search effort needed under this intent. Meanwhile, *P(click|hover)* (the probability of a result to be clicked given hovered [47]) is significantly lower in the *Procedure* intent than the others, indicating the user might skip more irrelevant results based on her preliminary judgments. It also suggests that the existing result list might not well satisfy the information need of *Procedure*.

Dwell Time. Although *Particular Case(s)* and *Learning* tasks involve similar numbers of queries and visited pages, the total task time is significantly longer in *Learning*. In particular, users spent more time on SERPs under the *Particular Case(s)* intent, while they spend more time on clicked case documents under the *Learning* intent. Compared among *Characterization*, *Penalty*, and *Procedure*, unsurprisingly, the *Penalty* tasks tend to take much more task time. Specifically, it took remarkably more time to examine the clicked results under the *Characterization* intent. Since the target information regarding the *Characterization* intent is usually broader in scope, users may need to read more contents in a case document to understand the entire case well.

Summary. Users' search behavior in legal case retrieval varies significantly with search intents regarding various aspects. Comparing *Particular Case(s)* and *Learning* intents, users seem more patient and spend more time examining the content of case documents in *Learning*. With the requirement for high precision and recall, users with *Particular Case(s)* intent might put more effort into exploring the SERPs. Among *Characterization*, *Penalty*, and *Procedure*, the *Penalty* tasks always involve the most search effort. Meanwhile, we observe that users with *Procedure* intent seem quite patient but less satisfied with the system results.

Just Accepted

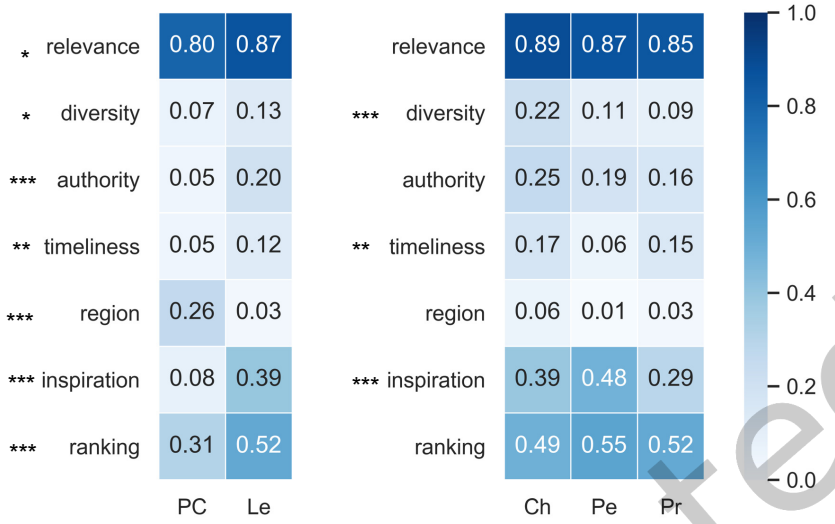


Fig. 4. The distribution of click reasons across different intents. The number in the grid denotes the proportion of the users that select the factor under the intent, correspondingly. "*/**/***" indicates the statistical significance at $p < 0.05/0.01/0.001$ level by one-way ANOVA, respectively.

Just Accepted

5.3 User Satisfaction under Different Intents

User satisfaction is a key concept in information retrieval systems, measuring the fulfillment of a user's information need [19, 54]. To answer **RQ3**, we first investigate how user satisfaction distributes under each intent and the influential factors according to users' explicit feedback. Furthermore, towards the evaluation of legal case retrieval, we attempt to measure user satisfaction with online metrics based on implicit signals.

5.3.1 Explicit Feedback. We observe significant differences in user satisfaction feedback across search intents. The average query satisfaction under each intent (i.e., *PC*, *Ch*, *Pe*, *Pr*) is 3.441, 3.207, 3.127, and 2.950, respectively ($p < 0.01$). Specifically, users perceive significantly higher satisfaction in the *Particular Case(s)* scenario than in the *Learning* ($p < 0.001$). Comparing the three categories within the *Learning*, the difference is mainly between *Procedure* and the others. Users seem not well satisfied in the *Procedure* context. In that case, the legal case retrieval systems need to put more effort into satisfying users' *Learning* tasks, and especially, attach due importance to the *Procedure* ones.

Further, we inspect the potential factors influencing user satisfaction under different search intents based on users' feedback on reasons for click-through (collected in *Step4)*. Following previous studies [11, 63], we consider the click as an important indicator of satisfaction. Firstly, no new factors outside of our seven options were proposed by the participants in our study. Therefore, we focus on the seven factors that we provided in the user study, i.e., *relevance*, *diversity*, *authority*, *timeliness*, *region*, *inspiration*, *ranking*. Figure 4 shows the distributions of these factors.

Across all search intents, *relevance* is always the main concern. However, beyond relevance, users pay attention to different aspects under different search intents. We observe that users may emphasize different aspects of a legal case document when search under the intents of Particular cases and Learning. Locality is more important when searching for particular cases, than we searching to satisfy an information need. On the contrary, when users search for *Learning*, they tend to care about other properties of the case contents, such as *authority*, *diversity*,

Table 7. Online Metrics and their descriptions.

Group	Metrics	Description
Click	UCTR	a binary variable indicating whether there was a click or not
	QCTR	the number of clicks
	MaxRR/MinRR/MeanRR	maximum/minimum/mean reciprocal ranks (RR) respectively
Dwell	SumClickDwell	the sum of click dwell time
	AvgClickDwell	the average of click dwell time
	QueryDwell	dwell time of the query session
	TimeToFirstClick	time delta between the start of the query and the first click
	TimeToLastClick	time delta between the start of the query and the last click

and *timeliness*. It is worth mentioning that *inspiration*, which means the result could inspire users to formulate better queries or find other cases, is emphasized under the *Learning* intent more often. We believe that the inspiration could help the user's exploratory search process. Moreover, the system ranking is more critical in the *Learning* tasks. Considering the larger result set and higher effort in examining results, we think users would rely more on the system rankings to identify better results. It also highlights the importance of optimizing top-ranked results in legal case retrieval, especially for the *Learning* tasks, although users might be more patient than in general Web search [47].

Significant differences mainly lie in *diversity*, *timeliness*, and *inspiration* when we compare among *Characterization*, *Penalty*, and *Procedure*. In particular, the *Characterization* intent requires a much higher level of result *diversity* than the others. The *inspiration* factor is more influential in the search intents regarding the substantive law, especially in the *Penalty* intent. Since users tend to put the most search effort into the *Penalty* tasks, the results with high inspiration would benefit the user's exploratory process. Meanwhile, results for the issues under the procedural law are usually within a more definite scope than those under the substantive law, which may be why users care less about result *diversity* and *inspiration* under the *Procedure* intent.

To sum up, users pay attention to different factors beyond relevance (e.g, *diversity*, *region*, *inspiration*, etc.) given different search intents. The results also shed light on the optimization directions for legal case search systems to promote user satisfaction under different search intents.

5.3.2 Implicit Signals. Evaluation plays an essential role in IR research, which measures how well the search system satisfies users' information needs. In contrast to offline evaluation metrics that rely on external relevance judgments, online metrics calculated based on behavioral logs (implicit signals) are cheaper and widely adopted in current search engines for system evaluation. Although the evaluation metrics designed for legal case search are still under investigation, this paper focuses on the performance of some common metrics generally applied to diverse search scenarios, taking user satisfaction as the "golden standard" [1, 10, 68]. To be specific, we conduct a correlation analysis to investigate how existing online metrics could measure user satisfaction, especially under different legal search intents. Following previous research [10, 68], we inspect the popular click-based and dwell-based metrics. Table 7 shows the online metrics that we use in this paper and their definitions. The Pearson's correlation coefficients between these online metrics and user satisfaction are shown in Table 8.

Click-base Metrics. UCTR and *QCTR* correlate significantly and positively with user satisfaction across all intent categories. Unlike the negative correlations in web search [10], users usually need to examine the case document to satisfy information needs, and thus more interactions with results are desired. Comparing among the search intents, the correlations with user satisfaction become weaker when the user's information need is relatively more specific (e.g., *Particular Case(s)* and *Penalty*). Similar trends can also be observed in metrics based

Just Accepted

Table 8. Pearson's correlation between online metric and user satisfaction under different intents. * indicates the correlation is significant at $p < 0.001$.

Group	Metric	PC	Ch	Pe	Pr
Click	UCTR	0.2815*	**0.4608***	0.2238*	0.3927*
	QCTR	0.2567*	**0.4447***	0.2943*	0.3581*
	MaxRR	0.2532*	**0.4324***	0.2300*	0.3681*
	MinRR	0.1475	**0.2990***	0.0818	0.2504*
	MeanRR	0.2040	**0.3820***	0.1572	0.3234*
Dwell	SumClickDwell	0.3162*	**0.4748***	0.3022*	0.3351*
	AvgClickDwell	0.2821*	**0.4362***	0.2539*	0.3155*
	QueryDwell	0.1903	0.2067	0.1679	**0.2875***
	TimeToFirstClick	0.0056	-0.0152	-0.2045	-0.0520
	TimeToLastClick	0.3312	0.1819	**0.3419***	0.1670

on click-through ranks. Specifically, *MinRR* and *MeanRR* can not well measure user satisfaction in the *Particular Case(s)* and *Penalty* scenarios. Only *MaxRR*, indicating the top rank of click, has significant correlations with user satisfaction under all search intents.

Dwell-based Metrics. *SumClickDwell* and *AvgClickDwell* have significant correlations with user satisfaction under all search intents. More time spent on examining case documents is a positive signal in legal case retrieval. However, *QueryDwell*, calculated based on the query's total dwell time, only correlates significantly with user satisfaction given the *Procedure* intent. Meanwhile, *TimeToLastClick* seem more suitable to measure user satisfaction under the intents that involve relatively specific information needs, such as *Penalty* and *Particular Case(s)* ($p = 0.009$).

In summary, online metrics demonstrate varying performances when used as indicators of user satisfaction. Given the diversity of search intents, it is essential to reconsider the extent to which a metric can accurately reflect user satisfaction and effectively evaluate the system.

6 APPLICATIONS

In this section, we attempt to apply the intent taxonomy to two critical downstream IR tasks to answer **RQ4** (*How can the taxonomy benefit downstream tasks in legal case retrieval*), including satisfaction prediction and result ranking.

6.1 Satisfaction Prediction

We attempt to predict user satisfaction with behavioral signals. In particular, we investigate the application of the intent taxonomy to this task from multiple perspectives. First, we inspect the performance of different behavioral signals in satisfaction prediction under different search intents. Second, we build an intent-aware model for satisfaction prediction.

6.1.1 Features. User behavior has been popularly utilized to predict satisfaction in varied search scenarios, such as Web search [21], product search [55], and image search [63]. However, there is limited research dedicated to constructing models for predicting user satisfaction in legal case retrieval. Referring to previous works [21, 55, 63] and preliminary analyses in the former sections, we extracted four groups of behavioral features (20 in total), as shown in Table 9. Features in the *Click*, *Hover*, and *Dwell* groups are the same as described in Section 5. In the

Just Accepted

Table 9. Behavioral features used in satisfaction prediction.

Feature Group	Feature Description	Numbers
Click	the number of clicks; the click-through rate; maximum/minimum/mean reciprocal ranks of clicks;	5
Hover	the number of hovers; the probability of being clicked given hovered; average of skipped results between hovers; maximum/minimum/mean ranks of hovers;	6
Dwell	dwell time on SERP/Landing Pages; time to first click; average of dwell time on hovered results; average of dwell time on clicked results;	5
Query	the length of query (in characters); the number of query terms; the ratio of unique terms; the number of visited pages;	4

Query group, we mainly utilized features that potentially reflect the overall complexity of this search through text statistics and browse pages. Note that we only used implicit signals (logged behavior) in this task and did not include any explicit feedback, considering that explicit feedback is rather expensive to collect in practice.

6.1.2 Experimental Settings. The behavioral dataset we used was collected in the user study as described in Section 5. Following previous research [54, 63], we mapped the 5-level satisfaction scale to a binary indicator (dissatisfied: 1&2&3, satisfied: 4&5) and treated satisfaction prediction as a binary classification task. Prediction performance was evaluated by AUC considering the imbalanced distribution of labels (dissatisfied: 470, satisfied: 345). We applied a gradient boosting decision tree model implemented by CatBoost [38], which can support both numerical and categorical features simultaneously and achieve great quality stably without parameter tuning. We considered two types of experimental settings, i.e., satisfaction prediction on the tasks of each intent and of all intents. Specifically, in the latter setting (denoted as "All Tasks" in Table 10, we compared the performance of intent-agnostic and intent-aware models. The intent-agnostic models are built based on the behavioral features listed in Table 9 and trained on the tasks of all intents. The intent-aware models added the intent category to the behavioral features and trained on the same data of the intent-agnostic models. Experiments were all conducted on 5-fold cross-validation.

6.1.3 Prediction Results. Results are as shown in Table 10. According to the prediction performance on "tasks per intent", we observe differences in the performance of behavioral features under different search intents. Specifically, the *Dwell* features achieve the best performance under the *Characterization* and *Penalty* tasks, while the *Query* features are more effective under the *Particular Case(s)* and *Procedure* intents. Furthermore, combining all kinds of features does not always lead to improvements. For instance, the combination of all features achieves the best performance only under the *Particular Case(s)* and *Characterization* intents. However, the combination may lead to a drop under the other intents, especially in the *Procedure* tasks. Meanwhile, comparing the prediction performance under different intents, user satisfaction under the *Procedure* intent seems the most difficult to model, which might need further effort to optimize. The results suggest that different types of behavioral signals should be utilized for satisfaction prediction when the search intent varies.

Just Accepted

Table 10. Satisfaction prediction performance measured by AUC. Results in boldface denote the best feature group and the best performance for each column.

	Tasks per Intent				**All Tasks**	
	PC	Ch	Pe	Pr	*Intent-agnostic*	*Intent-aware*
Click	0.6314	0.7135	0.6025	0.5901	0.6150	0.6366
Hover	0.6216	0.7145	0.6221	0.6261	0.6536	0.6523
Dwell	0.5893	**0.7255**	**0.6776**	0.6395	**0.6766**	**0.6918**
Query	**0.6409**	0.6831	0.5898	**0.6685**	0.5854	0.6187
All Features	**0.6996**	**0.7557**	0.6648	0.6294	0.6728	**0.7020**

According to the performance comparison on "all tasks", the intent-aware model performs better than the intent-agnostic model most of the time, given different features. In particular, when using the *Dwell* features and combination of features (*All Features*), which perform relatively better than other feature groups, the intent-aware methods consistently demonstrate significant improvements (t-test, $p < 0.05$) in performance. Specifically, the combination of behavioral features achieves the best performance with the search intent provided. The results suggest that involving search intent categories can contribute to improvements in satisfaction prediction performance.

6.2 Ranking

Ranking is a core task for IR. In this section, we exploited the widely adopted Learning to Rank (LTR) [9, 27, 31] and attempted to integrate the proposed intent taxonomy into this task.

6.2.1 Model. We follow the intent-aware ranking adaption framework [15] to integrate search intents with ranking models. To be specific, the probability that a result r satisfies the query q is calculated as,

$$P(r|q) = \sum_{i \in I} P(i|q)P(r|q,i) \tag{1}$$

, where I denotes the intent set. $P(r|q,i)$ is denotes the probability that the result r satisfies the query q under the intent i and is calculated by a intent-specific ranking module. Similar to [15], the intent-specific ranking module is an LTR model optimized for a specific intent. $P(i|q)$ denotes the probability of intent i given the query q. In our study, $P(i|q)$ works as an indicator ($P(i|q) \in \{0, 1\}$) indicates which intent the query belongs to. We acknowledge that this simplification would be a limitation since it dismissed the mixture of multiple intents. However, developing complicated intent-aware ranking models is beyond this paper's main concern, and we leave it for future work.

6.2.2 Experimental Settings. We utilized the sampled query logs with intent annotations as described in Section 4.1. We filtered out the data that were aggregated as *Multi* to avoid noise and the data in the *Others* or *Interest* categories due to the data sparsity. The clicked results were viewed as relevant, and the left were regarded as irrelevant. We filtered out the queries without any clicks. After filtering, we ended up with 525 search sessions under four intents, consisting of 1,177 queries. Case documents were downloaded. We extracted content-based features that have been commonly used in the LTR literature [31], including average term frequency (TF), average inverse document frequency (IDF), average TF-IDF, BM25 score, and cosine similarity based on TF-IDF vectors. All the models in the experiment used the same feature set. Regarding the learning algorithm, we employed three ranking algorithms: LambdaMART [62], RankBoost [13], and AdaRank [66]. These algorithms cover point-wise, pair-wise,

Just Accepted

Table 11. Comparison of intent-agnostic and intent-aware ranking models. Results in boldface denote the winner performance given each LTR model.

	NDCG@5	NDCG@10	NDCG@15	MAP
AdaRank	0.3811	0.4757	0.5361	0.3951
w/ intent-aware	**0.4444**	**0.5270**	**0.5801**	**0.4471**
improv.	+16.6%	+10.9%	+8.21%	+13.2%
LambdaMART	0.4936	0.5753	0.6105	0.4852
w/ intent-aware	**0.5329**	**0.6033**	**0.6354**	**0.5208**
improv.	+7.96%	+4.87%	+4.08%	+7.34%
RankBoost	0.5711	0.6446	0.6704	0.5648
w/ intent-aware	**0.5820**	**0.6524**	**0.6777**	**0.5737**
improv.	+1.91%	+1.21%	+1.09%	+1.58%

and list-wise ranking methodologies. For each ranking algorithm, we compared the performance between the intent-agnostic model and the intent-aware one. To be specific, the intent-agnostic model was trained on the queries of all intent categories. Under the intent-aware framework, the intent-specific module was trained based on the queries under the corresponding intent. The final ranking score was calculated according to formula (1). Both intent-agnostic and intent-aware models were tested on the same testing set, a mixture of queries under various intents. The algorithms were implemented by RankLib[9]. Parameters were set as default. We performed a five-fold cross-validation. In each cross-validation round, we used 10% of the train data as the validation set and optimized NDCG@10. The final performance was evaluated by NDCG@5, NDCG@10, NDCG@15, and MAP. Table 11 shows the average performance.

6.2.3 Results. As shown in Table 11, integrating search intents into ranking improves the performance of existing intent-agnostic ranking algorithms in legal case retrieval. The intent-aware models trained ranking models separately based on the training data with different intents. If user's needs and concepts of relevance do not vary across different intents in the developed intent taxonomy, the consideration of intents in the intent-aware models would only add noise to the training data, and make the final ranking models perform worse because the split of training data would make each model has less data for parameter optimization. However, as shown in Table 11, even with less training data for each intent ranking model, the intent-aware models still outperform the intent-agnostic models that trained with the full data. This demonstrates that users who submitted the queries with different intents under our intent taxonomy indeed have different needs and concepts of relevance. The benefits of intent information developed under our taxonomy are applicable to different ranking algorithms. Especially for some relatively weaker algorithms in this task (e.g., AdaRank and LambdaMART), the improvements in ranking performance are significant on all evaluation metrics (t-test, $p < 0.05$). Note that we are not intended to propose new ranking models for legal case retrieval in this study. Therefore, we utilized a simple but effective intent-aware adaption framework that could apply to varied ranking models. In conclusion, experimental results suggest the effectiveness of considering the intent taxonomy in the result ranking task.

6.3 Summary

In this section, we integrated the proposed intent taxonomy into two critical downstream IR tasks, i.e., satisfaction prediction and result ranking. In satisfaction prediction, we find that behavioral features play different roles

[9]https://sourceforge.net/p/lemur/wiki/RankLib/

Just Accepted

under different search intents. Moreover, involving search intents can improve the performance of satisfaction. In the ranking task, experimental results also suggest the effectiveness of the intent taxonomy by applying an intent-aware adaption.

7 DISCUSSIONS AND IMPLICATIONS

Understanding users' search intents is fundamental for search systems to satisfy users' information needs. Towards an in-depth investigation of search intents in legal case retrieval, this paper proposed a novel hierarchical intent taxonomy of legal case retrieval that integrates IR and legal classification theory. Regarding **RQ1** (*What are the types of user intent in legal case retrieval?*), user search intents can be categorized into five types: search for *Particular Case(s)*, *Characterization*, *Penalty*, *Procedure*, and *Interest*. According to the interviews and editorial studies, the proposed taxonomy has good coverage of the search intents in real-life search practice. Furthermore, the distribution of these intent categories are revealed based on the feedback collected in the verification process. The *Characterization* accounts for the largest proportion among all the intent categories. Meanwhile, the *Penalty* and *Procedure* are also worth research attention. Especially for the *Penalty* intent, all the interviewees have emphasized its importance in legal practice and, meanwhile, the difficulty in satisfying this intent.

Furthermore, towards modeling user behavior and satisfaction under different search intents (**RQ2** and **RQ3**), we conducted a laboratory user study involving 36 participants majoring in law. Implicit behavioral signals and explicit feedback were collected. Several interesting findings were made.

Regarding **RQ2** (*How does user search behavior change with search intents in legal case retrieval?*), we observe significant differences in user behavior under different search intents. In particular, we follow the hierarchical structure of the taxonomy and find differences when applying the criteria successively. Compared to the *Particular Case(s)* intent (divided by **Criterion 1**), users tend to be more patient and allocate more time to examine the landing page under the *Learning* intent. Comparing the intents classified by **Criterion 3** (i.e., *Characterization*, *Penalty*, and *Procedure*), *Penalty* involves the most effort. Meanwhile, we observe that users seem quite patient but less satisfied with the clicked results under the *Procedure* intent.

Regarding **RQ3** (*What are the differences in perception and measurement of user satisfaction under different search intents?*), search intents have significant influences on user satisfaction in multiple aspects. We observe that users are less satisfied under the *Learning* intent, especially under the *Procedure*. Although relevance is still the biggest concern in user satisfaction, users care about different factors (e.g., diversity, region, inspiration, ranking, etc.) given different search intents. With user satisfaction as the "golden standard", we find that the popularly adopted online metrics show distinct performance in measuring user satisfaction under different search intents in legal case retrieval. Our results suggest that the optimization and evaluation of search systems may also need to be adapted to different search intents in legal case retrieval.

Last but not least, we attempted to apply the intent taxonomy to other downstream tasks(e.g., satisfaction prediction and result ranking) to address **RQ4** (*How can the taxonomy benefit downstream tasks in legal case retrieval?*). Experimental results demonstrate the benefits of the intent taxonomy in legal case retrieval.

Implications. This work provides insight into user intents in the scenario of legal case retrieval. It provides a fundamental research contribution to related studies in legal case retrieval, such as relevance criteria, ranking strategies, and evaluation design. While our taxonomy was originally developed and validated within the Chinese legal system, it provides a solid foundation that can inspire the development of similar taxonomies in other legal systems. The underlying principles and categorization framework can serve as a starting point for researchers and practitioners working in different jurisdictions.

Extensive experimental results based on various sources suggest the significance of considering different types of search intents in legal case retrieval. Recent research efforts [34, 46, 47] on legal case retrieval are mainly concerned with the *Characterization* tasks, which accounts for the most significant proportion of search intents

Just Accepted

in legal case retrieval according to our study. However, we argue that other intent types, such as *Penalty* and *Procedure*, are also worth investigation and optimization in the meantime, which still lack due research attention. Moreover, our study has revealed remarkable influences of search intents on various components of an IR system, such as search behavior, user satisfaction, system evaluation, and result ranking. Therefore, our findings suggest that the methodologies in legal case retrieval should be adjusted with various search intents instead of merely similar case matching. This work provides promising implications for the development of legal case retrieval systems to better satisfy users' diverse information needs in practice, such as developing intent-aware ranking strategies and evaluation metrics.

8 CONCLUSION AND FUTURE WORK

In this paper, we present a novel intent taxonomy for legal case retrieval. To our best knowledge, it is the first taxonomy that categorizes users' search intents in legal case retrieval. The taxonomy was built based on various resources and further evaluated extensively via interviews and editorial user studies. Furthermore, based on an additional laboratory user study, we discovered significant differences in search behavior and satisfaction under different search intents of legal case retrieval. Finally, we applied the intent taxonomy to two essential tasks in legal case retrieval and demonstrated its implications.

We acknowledge some potential limitations of this work. The experiments in this paper were mainly conducted based on the Chinese law system, e.g., user studies and query logs, though the taxonomy is designed to be generally applicable across different law systems. The impacts regarding other law systems may need further studying. As for the user study design in Section 5, the number of participants and tasks is limited as in most user studies, especially for the search scenarios involving domain knowledge [47, 69]. The *Interest* category still lacks an in-depth inspect limited by the user study environment and data sparsity in query logs. In Section 6, traditional models were utilized and the way of integrating search intents seemed straightforward, since our primary concern is the influence of intents on these tasks. More complicated models are out of scope and left for future work.

As for future work, we will work on developing intent-aware mechanisms for legal case retrieval. For instance, we plan to construct benchmarks with different search intents involved and design intent-aware evaluation metrics. Besides, intent-aware ranking strategies are worth investigating to satisfy diverse information needs better in legal case retrieval. To resolve the user study's limitations, a larger-scale field study will be a promising supplementary for future work.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China (Grant No. 61732008, 62002194) and Tsinghua University Guoqiang Research Institute.

Just Accepted

REFERENCES

[1] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The relationship between IR effectiveness measures and user satisfaction. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 773–774.

[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] Steven M Barkan, Barbara Bintliff, and Mary Whisner. 2015. Fundamentals of legal research. (2015).

[4] 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.

[5] Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W Bruce Croft, and Mark Sanderson. 2022. A Non-Factoid Question-Answering Taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1196–1207.

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

[7] B Barla Cambazoglu, Leila Tavakoli, Falk Scholer, Mark Sanderson, and Bruce Croft. 2021. An intent taxonomy for questions asked in web search. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 85–94.

[8] Olivier Chapelle, Shihao Ji, Ciya Liao, Emre Velipasaoglu, Larry Lai, and Su-Lin Wu. 2011. Intent-based diversification of web search results: metrics and algorithms. *Information Retrieval* 14, 6 (2011), 572–592.

[9] Jia Chen, Haitao Li, Weihang Su, Qingyao Ai, and Yiqun Liu. 2023. THUIR at WSDM Cup 2023 Task 1: Unbiased Learning to Rank. arXiv:2304.12650 [cs.IR]

[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] Ovidiu Dan and Brian D Davison. 2016. Measuring and predicting search engine users' satisfaction. *ACM Computing Surveys (CSUR)* 49, 1 (2016), 1–35.

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

[13] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *Journal of machine learning research* 4, Nov (2003), 933–969.

[14] 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.

[15] Rafael Glater, Rodrygo LT Santos, and Nivio Ziviani. 2017. Intent-aware semantic query annotation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 485–494.

[16] 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.

[17] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.

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

[19] Diane Kelly et al. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224.

[20] Eugene Kharitonov, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. 2013. Intent models for contextualising and diversifying query suggestions. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2303–2308.

[21] Youngho Kim, Ahmed Hassan, Ryen W White, and Imed Zitouni. 2014. Comparing client and server dwell time estimates for click-level satisfaction prediction. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 895–898.

[22] Michel CA Klein, Wouter Van Steenbergen, Elisabeth M Uijttenbroek, Arno R Lodder, and Frank van Harmelen. 2006. Thesaurus-based Retrieval of Case Law. *Frontiers in Artificial Intelligence and Applications* 152 (2006), 61.

[23] Christoph Kofler, Martha Larson, and Alan Hanjalic. 2016. User intent in multimedia search: a survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)* 49, 2 (2016), 1–37.

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

[25] 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.

[26] Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023. SAILER: Structure-aware Pre-trained Language Model for Legal Case Retrieval. arXiv:2304.11370 [cs.IR]

[27] Haitao Li, Jia Chen, Weihang Su, Qingyao Ai, and Yiqun Liu. 2023. Towards Better Web Search Performance: Pre-training, Fine-tuning and Learning to Rank. arXiv:2303.04710 [cs.IR]

Just Accepted

[28] Haitao Li, Weihang Su, Changyue Wang, Yueyue Wu, Qingyao Ai, and Yiqun Liu. 2023. THUIR@COLIEE 2023: Incorporating Structural Knowledge into Pre-trained Language Models for Legal Case Retrieval. arXiv:2305.06812 [cs.IR]
[29] Haitao Li, Changyue Wang, Weihang Su, Yueyue Wu, Qingyao Ai, and Yiqun Liu. 2023. THUIR@COLIEE 2023: More Parameters and Legal Knowledge for Legal Case Entailment. arXiv:2305.06817 [cs.CL]
[30] Bulou Liu, Yueyue Wu, Yiqun Liu, Fan Zhang, Yunqiu Shao, Chenliang Li, Min Zhang, and Shaoping Ma. 2021. Conversational vs Traditional: Comparing Search Behavior and Outcome in Legal Case Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1622–1626.
[31] Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.
[32] Daniel Locke and Guido Zuccon. 2022. Case law retrieval: problems, methods, challenges and evaluations in the last 20 years. arXiv:2202.07209 [cs.IR]
[33] Mathias Lux, Christoph Kofler, and Oge Marques. 2010. A classification scheme for user intentions in image search. In *CHI'10 Extended Abstracts on human factors in computing systems*. 3913–3918.
[34] Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. LeCaRD: A Legal Case Retrieval Dataset for Chinese Law System. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. 2342–2348.
[35] Arpan Mandal, Kripabandhu Ghosh, Saptarshi Ghosh, and Sekhar Mandal. 2021. Unsupervised approaches for measuring textual similarity between legal court case reports. *Artificial Intelligence and Law* (2021), 1–35.
[36] Rishabh Mehrotra, Mounia Lalmas, Doug Kenney, Thomas Lim-Meng, and Golli Hashemian. 2019. Jointly leveraging intent and interaction signals to predict user satisfaction with slate recommendations. In *The World Wide Web Conference*. 1256–1267.
[37] Marie-Francine Moens. 2001. Innovative techniques for legal text retrieval. *Artificial Intelligence and Law* 9 (2001), 29–57.
[38] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* 31 (2018).
[39] 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.
[40] Daniel E Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*. 13–19.
[41] Juline Rossi and Evangelos Kanoulas. 2019. Legal information retrieval with generalized language models. *Proceedings of the 6th Competition on Legal Information Extraction/Entailment. COLIEE* (2019).
[42] Jaromir Savelka and Kevin Ashley. 2021. Discovering Explanatory Sentences in Legal Case Decisions Using Pre-trained Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 4273–4283. https://doi.org/10.18653/v1/2021.findings-emnlp.361
[43] Jaromír Šavelka and Kevin D Ashley. 2022. Legal information retrieval for understanding statutory terms. *Artificial Intelligence and Law* (2022), 1–45.
[44] Jaromir Savelka, Huihui Xu, and Kevin D Ashley. 2019. Improving sentence retrieval from case law for statutory interpretation. In *Proceedings of the seventeenth international conference on artificial intelligence and law*. 113–122.
[45] 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*. 3501–3507.
[46] Yunqiu Shao, Yueyue Wu, Yiqun Liu, Jiaxin Mao, and Shaoping Ma. 2022. Understanding Relevance Judgments in Legal Case Retrieval. *ACM Transactions on Information Systems (TOIS)* (oct 2022). https://doi.org/10.1145/3569929 Just Accepted.
[47] 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* (Virtual Event, Canada) *(SIGIR '21)*. 962–972.
[48] Emily Sherwin. 2009. Legal taxonomy. *Legal Theory* 15, 1 (2009), 25–54.
[49] Olga Shulayeva, Advaith Siddharthan, and Adam Wyner. 2017. Recognizing cited facts and principles in legal judgements. *Artificial Intelligence and Law* 25, 1 (2017), 107–126.
[50] Mathias M Siems. 2016. Varieties of legal systems: towards a new global taxonomy. *Journal of Institutional Economics* 12, 3 (2016), 579–602.
[51] Parikshit Sondhi, Mohit Sharma, Pranam Kolari, and ChengXiang Zhai. 2018. A taxonomy of queries for e-commerce search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1245–1248.
[52] Holger Spamann, Lars Klöhn, Christophe Jamin, Vikramaditya Khanna, John Zhuang Liu, Pavan Mamidi, Alexander Morell, and Ivan Reidel. 2021. Judges in the Lab: No Precedent Effects, No Common/Civil Law Differences. *Journal of Legal Analysis* 13, 1 (03 2021), 110–126. https://doi.org/10.1093/jla/laaa008 arXiv:https://academic.oup.com/jla/article-pdf/13/1/110/41986764/laaa008.pdf
[53] James A Sprowl. 1975. The Westlaw System-A Different Approach to Computer-Assisted Legal Research. *Jurimetrics J.* 16 (1975), 142.
[54] Louise T Su. 1992. Evaluation measures for interactive information retrieval. *Information Processing & Management* 28, 4 (1992), 503–516.

Just Accepted

[55] Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. 2018. User intent, behaviour, and perceived satisfaction in product search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 547–555.
[56] Paul Thomas, Bodo Billerbeck, Nick Craswell, and Ryen W White. 2019. Investigating searchers' mental models to inform search explanations. *ACM Transactions on Information Systems (TOIS)* 38, 1 (2019), 1–25.
[57] 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 *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. 275–282.
[58] Yaqing Wang, Song Wang, Yanyan Li, and Dejing Dou. 2022. Recognizing Medical Search Query Intent by Few-shot Learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 502–512.
[59] Hannes Westermann, Jaromir Savelka, and Karim Benyekhlef. 2020. Paragraph similarity scoring and fine-tuned BERT for legal information retrieval and entailment. In *JSAI International Symposium on Artificial Intelligence*. Springer, 269–285.
[60] Ryen W White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. 2013. Enhancing personalized search by mining and modeling task behavior. In *Proceedings of the 22nd international conference on World Wide Web*. 1411–1420.
[61] Glanville Williams. 1982. Substantive and Adjectival Law. In *Learning the Law*. Law Book Company, Limited, 19–23.
[62] Qiang Wu, Chris JC Burges, Krysta M Svore, and Jianfeng Gao. 2008. *Ranking, boosting, and model adaptation*. Technical Report. Technical report, Microsoft Research.
[63] Zhijing Wu, Yiqun Liu, Qianfan Zhang, Kailu Wu, Min Zhang, and Shaoping Ma. 2019. The influence of image search intents on user behavior and satisfaction. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 645–653.
[64] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, et al. 2019. Cail2019-scm: A dataset of similar case matching in legal domain. *arXiv preprint arXiv:1911.08962* (2019).
[65] Xiaohui Xie, Yiqun Liu, Maarten De Rijke, Jiyin He, Min Zhang, and Shaoping Ma. 2018. Why people search for images using web search engines. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 655–663.
[66] Jun Xu and Hang Li. 2007. Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 391–398.
[67] Weijie Yu, Zhongxiang Sun, Jun Xu, Zhenhua Dong, Xu Chen, Hongteng Xu, and Ji-Rong Wen. 2022. Explainable Legal Case Matching via Inverse Optimal Transport-based Rationale Extraction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 657–668.
[68] Fan Zhang, Ke Zhou, Yunqiu Shao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. How Well do Offline and Online Evaluation Metrics Measure User Satisfaction in Web Image Search?. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 615–624.
[69] 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.

Just Accepted