Understanding Relevance Judgments in Legal Case Retrieval

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

Legal case retrieval, which aims to retrieve relevant cases given a query case, has drawn increasing research attention in recent years. While much research has worked on developing automatic retrieval models, how to characterize relevance in this specialized information retrieval (IR) task is still an open question. Towards an in-depth understanding of relevance judgments, we conduct a laboratory user study that involves 72 participants of different domain expertise. In the user study, we collect the relevance score along with detailed explanations for the relevance judgment and various measures of the judgment process. From the collected data, we observe that both the subjective (e.g., domain expertise) and objective (e.g., query/case property) factors influence the relevance judgment process. By investigating the collected user explanations, we identify task-specific patterns of user attention distribution and re-think the criteria for relevance judgments. Moreover, we investigate the similarity in attention distribution between models and users. Further, we propose a two-stage framework that utilizes user attention to improve relevance estimation for legal case retrieval. Our study sheds light on understanding relevance judgments in legal case retrieval and provides implications for improving the design of corresponding retrieval systems.

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

Additional Key Words and Phrases: Legal case retrieval, relevance judgment, user study

ACM Reference format:
Yunqiu Shao, Yueyue Wu, Yiqun Liu, Jiaxin Mao, and Shaoping Ma. 2023. Understanding Relevance Judgments in Legal Case Retrieval. *ACM Trans. Inf. Syst.* 41, 3, Article 76 (February 2023), 32 pages.
https://doi.org/10.1145/3569929

This work is supported by the Natural Science Foundation of China (Grant No. 61732008, 62002194) and Tsinghua University Guoqiang Research Institute.
Authors' addresses: Y. Shao, Y. Wu, Y. Liu (corresponding author), and S. Ma, Department of Computer Science and Technology, Institute for Internet Judiciary, Tsinghua University, Zhongguancun Laboratory, Quan Cheng Laboratory, Beijing, China; emails: shaoyq18@mails.tsinghua.edu.cn, wuyueyue1600@gmail.com, {yiqunliu, msp}@tsinghua.edu.cn. J. Mao, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China; email: maojiaxin@gmail.com.

1046-8188/2023/02-ART76
https://doi.org/10.1145/3569929

1 INTRODUCTION

Legal case retrieval is a specialized **Information Retrieval** (IR) task, which aims to retrieve relevant cases given a query case. It is of vital importance in pursuing legal justice in various legal systems. In common law, precedents are fundamental for legal reasoning, following the doctrine of *stare decisis*. Meanwhile, although prior cases are not always cited directly in some other law systems (e.g., Germany [19], China), they are still critical for supporting the decision-making process. In recent years, China has established a system of similar case retrieval[1] and continuously expanded the scope of compulsory retrieval[2] for consistency in legal decisions. With the rapid growth of digitalized case documents, how to identify relevant cases effectively has drawn increasing research attention in both legal and IR communities. In recent years, several benchmark datasets have been constructed, such as COLIEE [34], AILA [5], and LeCaRD [27]. They provide binary or graded relevance labels for training and evaluating legal case retrieval models [43, 56]. However, we still lack a solid understanding of *relevance* in legal case retrieval, especially how users make relevance judgments in this scenario, which may hinder future progress in this area.

Relevance is a key notion in IR. Generally, *relevant information* is defined as information that pertains to the matter of the problem at hand [40]. It has different manifestations, including the relevance calculated by algorithms of IR systems (i.e., system relevance) and the relevance that a user assesses (i.e., user relevance) [50]. The IR testing is generally based on the comparison between the system relevance and the user relevance, taking the user relevance as the gold standard [10, 40]. However, things become a bit more complex concerning legal case retrieval. As a specific task oriented to judicial applications, relevance judgments in legal case retrieval should follow the legal standards, in other words, consider validity in the legal domain. For instance, the relevance assessed by the user might not agree with the authoritative legal rule, and the corresponding result could not satisfy the information need in legal practice consequently. Table 1 gives an example of considering domain validity for making relevance judgments. According to the legal rules, the "theft" case is more relevant than the "credit card fraud" to the query case, where the defendant used a stolen credit card. Therefore, the domain validity should also be considered when discussing user relevance in legal case retrieval, which has not received due attention.

In the research line around relevance in IR, understanding how users determine relevance is essential for investigating the concept of relevance [16, 40]. In the legal domain, there also exist some user studies specific to the e-discovery task [32] to investigate the factors that affect user relevance judgments [9, 53, 54]. However, the task definition of e-discovery [32] differs from that of legal case retrieval. The e-discovery task is to retrieve various "electronically stored information" (e.g., letters, e-mails) regarding a request in legal proceedings, while the legal case retrieval task aims to identify the relevant legal cases that support the decision process of the given query case. Therefore, the definitions of relevance in the two tasks are different correspondingly. As far as we know, existing research efforts in legal case retrieval are mainly put into developing retrieval models (i.e., system relevance) and some theoretical discussions [45, 48], lacking an empirical and quantitative investigation of the user side. Concerning user relevance in legal case retrieval, we propose the first two research questions in this article:

- **RQ1**: What factors will affect the process of making relevance judgments in legal case retrieval?
- **RQ2**: How do users allocate their attention when making relevance judgments?

[1] http://english.court.gov.cn/2020-08/13/content_37538734.htm

[2] The original text: https://www.court.gov.cn/fabu-xiangqing-334151.html and an explanation in English: https://www.lawinfochina.com/Search/DisplayInfo.aspx?id=33217&lib=news.

Table 1. An Example of Considering Domain Validity in Determining the Case Relevance

Query Case	*Description*: ... On March 18, 2020, the defendant A had stolen his roommate's credit card and consumed 50,000 yuan in a shopping mall ...
Candidate_1 (NR)	*Description*: ... On April 5, 2019, the defendant B got his friend's credit card by claiming he could increase the credit limit for free and used it for luxury consumption of 150,000 yuan ... *Judgment*: Crime of credit card fraud
Candidate_2 (R)	*Description*: ... On May 3, 2018, the defendant C has stolen 3 wallets at the train station, including 7 credit cards and 3,500 yuan in cash ... *Judgment*: Crime of theft
Analysis: Concerning the query, "Candidate_2," rather than "Candidate_1," is a relevant case, although "Candidate_1" seems more similar to the query case description (e.g., credit card, consumption). From a legal perspective, using a stolen credit card may be suspected of theft instead of credit card fraud according to the Criminal Law.	

Given the manifestations of relevance, we further put efforts into interpreting the gap between user relevance and system relevance in legal case retrieval. Specifically, our third research question is:

- **RQ3**: Do retrieval models pay attention to similar contents with users? Can we utilize the user's attention to improve the model for legal case retrieval?

To address these research questions, we conducted a laboratory user study that involved 72 participants with different levels of domain expertise. Beyond relevance judgments, details of the decision-making process were collected, including user feedback, highlights, and multi-aspect reasons. With the collected data, we systematically investigated the process of making relevance judgments and the corresponding reasons. Furthermore, we investigated system relevance in an explainable way by comparing the models' attention with users and thus proposed a two-stage framework to improve the retrieval models in legal case retrieval. With a better understanding of relevance judgments in legal case retrieval, our study provides implications for developing better case retrieval systems in legal practice.

2 RELATED WORK

2.1 Legal Case Retrieval

Finding relevant materials is a fundamental component of legal practice, and thus legal IR is always an active research topic in the legal and IR communities [4]. Extensive joint efforts have been put into developing the professional systems [2] to support the legal IR practice since the last century, such as Westlaw [14]. Centered on the specific e-discovery task [32], which refers to the process of one party's discovering evidence in the form of electronically stored information held by another party, a Legal Track was added to the TREC in 2006. The dataset constructed in the Legal Track supported further research regarding e-discovery, such as retrieval methods [55, 59] and relevance judgments [9, 54]. Unlike e-discovery, legal case retrieval aims to identify relevant cases decided in courts of law, which are primary legal materials along with statutes, to support the decision-making process. In recent years, there have been a number of benchmark datasets centered on the legal case retrieval task, such as COLIEE [34], AILA [5], and LeCaRD [27]. Among them, the datasets of COLIEE and AILA were constructed based on the common law systems in Canada and India, respectively. LeCaRD was constructed based on the Chinese law system. Given these benchmarks, the development of retrieval models for legal case retrieval was promoted, especially the NLP-based semantic matching methods [29, 37, 43, 47, 56]. However, the relevance provided in

the benchmarks [5, 27, 34] was in the form of binary or graded scales without further explanations, leaving the judgment-making process under-investigated. Meanwhile, some works have argued the importance of understanding the realistic interactions between users and systems. For example, Shao et al. [44] investigated users' search behavior in legal case retrieval and characterized it as an exploratory search process. Liu et al. [26] investigated the application of conversational search in this task. Different from the existing user studies [26, 44], which worked on the general search process (e.g., querying and clicking), we make a more focused and fine-grained investigation of the relevance judgments in legal case retrieval.

2.2 Relevance

Relevance is a key concept in information science, especially in IR. Generally, it measures the connection between given information and a given context problem at hand, which is an intensely human notion [40]. Much research has been done around the concept of relevance. It has different manifestations [7, 11, 39, 50], such as *systemic or algorithmic relevance*, *topical relevance or subjective relevance*, *cognitive relevance or pertinence*, *usefulness or situational relevance*, and *effective relevance* [40]. In particular, the classic IR model distinguished between *system relevance* and *user relevance,* and the IR testing is always based on the comparison between system and user relevance, such as the well-known Cranfield framework [10]. Under this testing framework, user assessment of relevance is considered as the gold standard for evaluating system effectiveness. A variety of studies further investigated how humans determine relevance via experimental studies on relevance behavior [16]. In the earlier research line, empirical studies [3, 28, 42, 46] were conducted to work on the criteria and clues for relevance assessments via interviewing users and coding responses into distinct categories. Moreover, eye-tracking [6, 18, 24] and brain imaging [1] have also been applied to investigating the cognitive process of relevance. For instance, through users' eye movements and annotated texts, Li et al. [24] investigated reading attention distribution on the result document during relevance judgment in web search and observed several reading behavior biases, e.g., position bias. Focused on the scenario of non-factoid QA, Bolotova et al. [6] inspected the attention distributions of neural networks and people.

One of the challenges in legal IR lies in the definition of legal relevance, which is beyond the topical relevance in web search and is complicated. Both theoretical and empirical studies have been conducted concerning legal relevance. From the theoretical view, Opijnen and Santos [48] argued that the role of the legal domain is essential along with the classical user-system interplay. In particular, they further discussed relevance in six concrete dimensions, including *algorithmic or system relevance*, *topical relevance*, *bibliographic relevance*, *cognitive relevance*, *situational relevance or utility*, and *domain relevance*. As for the empirical studies, there have been several experimental studies on users' relevance judgments concerning the e-discovery task, based on the test collections of TREC legal Track [32]. For instance, Chu [9] ranked factors that might affect relevance judgments from six categories according to the votes of nine participants without law backgrounds. The process of relevance judgments (e.g., accuracy, agreement, and perceived difficulty) in e-discovery was inspected through a user study that involved four law students and four information science students [53, 54]. It is worth noting that legal case retrieval differs from e-discovery in various aspects, such as task definition and retrieved targets, and thus relevance concepts should also vary. Regrading case retrieval, a relevant case should support the decision process [34, 44, 45]. Sutton [45] made a theoretical analysis of the role of attorney mental models in determining case relevance under the United States law system. Note that Sutton's research [45] differs from this article in multiple aspects. The concept of relevance in Sutton's [45] was considered dynamic and personalized. He mainly discussed the general information-seeking and evaluation process of attorneys searching the corpus of case law. The attorney mental models were presented in a high level of abstraction,

Table 2. The Selected Causes of Action for Tasks in the User Study

Category	Causes	Companion Causes
Endangering Public Security	Crime of Dangerous Driving	Crime of Disrupting Public Service
	Crime of Traffic Accident	Crime of Dangerous Driving
Encroaching on Property	Crime of Theft	Crime of Cover-up or Concealment of Crime-related Income and Proceeds, Crime of Duty Encroachment
	Crime of Fraud	Crime of Contract Fraud
	Crime of Extortion	Crime of False Imprisonment
Infringing Upon the Rights of the Person	Crime of Intentional Injury	Crime of Affray
	Crime of False Imprisonment	Crime of Robbery
Disrupting the Order of Social Administration	Crime of Disrupting Public Service	Crime of Intentional Injury, Crime of False Imprisonment
Graft and Bribery	Crime of Accepting Bribes	–
	Crime of Accepting Bribes as Non-official Servant	–

i.e., *base-level modeling*, *context-sensitive exploration*, and *disambiguation*, leaving a gap from application in the practical retrieval systems. However, the relevance discussed in this article is a more objective concept as applied in legal case retrieval benchmarks, and personalization is out of the scope. Instead of the general information-seeking process, we focus on the process of making relevance judgments between a query and a candidate case and conduct an empirical and quantitative analysis.

In this article, referring to the classic IR model [40], we investigate user relevance and system relevance in legal case retrieval, considering the domain validity (i.e., correctness) simultaneously. Focused on the specific but fundamental retrieval task in legal practice, the conclusions in our study can benefit a series of related works in legal case retrieval, including the construction of datasets and development of retrieval models.

3 USER STUDY

3.1 Tasks

Generally speaking, the relevance judgment task in our study is to determine the relevance of the candidate case, given a query case. Our tasks were constructed based on LeCaRD [27] since it is the largest public dataset for legal case retrieval for the Chinese law system. LeCaRD [27] consisted of 107 query cases and about 43,000 candidate cases in total, among which the top 30 candidate cases of each query case were annotated in four-level relevance scales. The query and candidate cases in LeCaRD [27] were adopted from the criminal cases published by the Supreme People's Court of China.[3] In particular, the query cases of LeCaRD [27] covered the top 20 frequent criminal charges.

We carefully selected 16 query cases from the 107 cases to design tasks in our user study. First, we attempted to cover the main categories of crimes in the Chinese law system. To be specific, we selected 10 of the charges (also labeled as "Cause of Action" or "Cause" for short) from LeCaRD (20 in total). As shown in Table 2, the selected causes involved the five main categories of crimes. Second, we wanted to involve diverse complexity in the selected query cases. Our study utilized the number of causes involved in the query case to control the query case complexity and considered it an independent variable. Based on the number of causes, the query cases could be divided into two groups, i.e., the **query with single cause** (**QSC**) and the **query with multiple causes** (**QMC**).

[3] https://wenshu.court.gov.cn/.

Table 3. Relevance Scales and Instructions in the User Study

Relevance Scale	Instruction
1 (Irrelevant)	Neither key circumstances nor key elements are relevant.
2 (Somewhat relevant)	Key circumstances are relevant but key elements are not.
3 (Fairly relevant)	Key elements are relevant but key circumstances are not.
4 (Highly relevant)	Both key circumstances and key elements are relevant.

The definitions of "key circumstance" and "key element" are the same as those in Table 5.

Note that almost every cause in LeCaRD contains query cases that involve either only one or multiple causes. Table 2 shows the selected causes and their companion causes in the QMC tasks. Notably, there were no QMC tasks under the "Crime of Accepting Bribes" and the "Crime of Accepting Bribes as Non-official Servant" because of the lack in LeCaRD. Meanwhile, because the QSC tasks under the "Crime of Dangerous Driving" and the "Crime of Traffic Accident" in LeCaRD were much easier than other selected cases, we dismissed them in our study. In total, we obtained eight QSC and eight QMC query cases, respectively. Last but not least, only the query case description was shown to the participant to simulate the realistic search scenario, consistent with the settings of previous work [34, 44].

We applied the four-level relevance scales referring to those of LeCaRD, as shown in Table 3. The instructions for relevance assessments were defined based on the "critical factor," which means factors significant to the constitutive elements of crime, including "key element" and "key circumstance." As shown in Table 5, the "key element" is defined as "the constitutive element of crime, which is the abstraction of key circumstances," and the "key circumstance" is defined as "the fact which is significant to the conviction and sentencing." More specifically, Table 4 provides an example[4] for key elements and key circumstances.

Then the candidate cases were selected as follows. One candidate case of each relevance level was selected for each query case in our study. We first selected the candidate case of the corresponding relevance score from LeCaRD. In particular, only the cases in which at least two assessors of LeCaRD agreed on the label were considered. If there were cases in which all the three LeCaRD assessors agreed, we selected randomly among them. Otherwise, we selected randomly from the left cases, i.e., those in which two assessors agreed on the relevance label. Then, if LeCaRD did not include any case of that relevance score, we found it manually from China judgments Online.[5] Generally, we constructed $16 \times 4 = 64$ query-document pairs (i.e., tasks).

3.2 Participants

We recruited 72 participants (22 males, 50 females) via universities' online forums and social networks. They were all native Chinese speakers. In particular, they were divided into three groups based on their domain expertise, i.e., the **criminal user** (**CU**), the **non-criminal user** (**NCU**), and the **non-legal user** (**NLU**). Each group contained 24 participants. To be specific, participants in the criminal or the non-criminal group all majored in law. They were graduate students in law school and qualified in legal practice.[6] However, the two groups differed in legal specialties. The criminal group majored in criminal law, while the non-criminal group majored in other specialties apart from criminal law, such as civil law, administrative law, etc. The non-legal group was composed of students without law backgrounds. Diverse majors were included in this group, including engineering, science, arts, literature, and so forth. Since the tasks were all criminal cases,

[4]This example along with referred article was also provided in the instructions for participants.
[5]https://wenshu.court.gov.cn/.
[6]They had passed the "National Uniform Legal Profession Qualification Examination."

Table 4. Examples for "Key Element" and "Key Circumstance"

Case (Partial)	This court believes that the defendant Zhang Dong in the original trial, together with Liu Bao and Li Nan, beat others at will, causing minor injuries to others at level two, and his actions disrupted social order, constituted the crime of picking and provoking trouble, and were a joint crime. ··· After investigation, the defendant Zhang Dong, together with Liu Bao and Li Nan, had accidental conflicts with the victim Shi, and beat the victim Shi without reason, causing him minor injuries. His behavior was consistent with the crime of quarreling and provoking quarrels. ···
Key Circumstance	Accidental conflicts; Beat without reason; Minor injuries at level two
Key Element	Assault others at will; Provoking troubles; Cause minor injuries to others; Disrupt social order
Reference (Article)	*Criminal Law Article 293: Anyone who commits one of the following acts of provoking troubles and disrupting social order shall be sentenced to fixed-term imprisonment of not more than five years, criminal detention or surveillance: (1) Assault others at will, with atrocious circumstances; (2) Chase or intercept, Insulting, intimidating others, and the circumstances are bad; (3) Forcing or arbitrarily destroying or occupying public or private property, the circumstances are serious; (4) Making disturbances in public places, causing serious disorder in public places. Anyone who gathers others to perform the acts mentioned in the preceding paragraph several times and seriously disrupts social order shall be sentenced to fixed-term imprisonment of not less than five years but not more than ten years, and may also be fined.*

Given a part of case documents, the corresponding key elements and key circumstances are listed as follows. The corresponding criminal law article is provided as reference in this example to help understand the key elements.

the criminal group was considered to be of the highest domain expertise in our study, followed by the non-criminal group and the non-legal group successively.

3.3 Procedure

The procedure of our user study is as illustrated in Figure 1. Before starting the experiment, the participants were provided with a detailed guideline, which included the study description, the definition of legal concepts (Table 5) used in the study, and the instruction and example for each relevance level. Participants of all domain expertise used the same guideline in our study. Specifically, the participants were instructed to assume that they are dealing with the given query case and need to decide the relevance level of the candidate case, measuring how well it can support the decision process of the query case. After signing the informed consent, they completed a warm-up task to get familiar with the experimental settings.

S1: Query Case Reading and Pre-task Questionnaire. At the beginning of each task, the participant read through the query case description and then filled in a pre-task questionnaire to report his or her expected task difficulty on a 5-point Likert-type scale (1: not at all, 5: very).

S2: Candidate Case Reading and Relevance Assessment. The participant was directed to the relevance assessment page, where a candidate case document was provided, followed by a relevance score form on the 4-point scale. The participant needed to decide how relevant the provided candidate case was to the query case. The participant could refer to the query case page at any time via the link on the top. The platform recorded the dwell time on the relevance assessment page as the time of making relevance judgments.

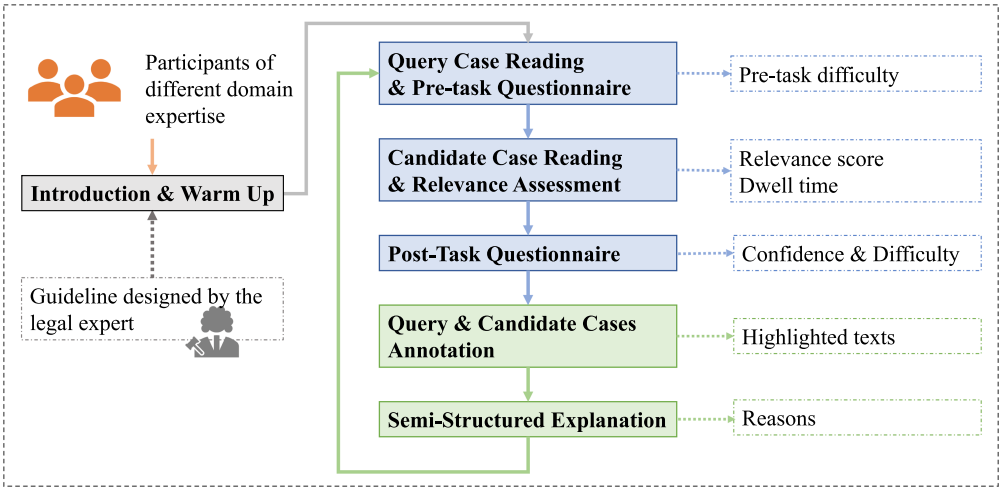


Fig. 1. The user study procedure.

Table 5. The Framework of Reasons in the User Study

Type	Aspect	Description
Required	Key Element	The constitutive element of crime, which is the abstraction of key circumstances.
	Key Circumstance	The fact that is significant to the conviction and sentencing.
Optional	Issue	A legal question, at the foundation of a case and requiring a court's decision, includes issue of law and issue of fact.
	Cause of Action	The legal generalization of key issue.
	Application of Law	The specific clause applied in the case.
	Other	Not belonging to any of the above.

S3: Post-task Questionnaire. Once submitting the relevance score, the participant was directed to the post-task questionnaire to report his or her perceived task difficulty and confidence while making the relevance judgment on a 5-point scale (1: not at all, 5: very).

S4: Query and Candidate Case Annotation. Then the participant was instructed to recall the process of making the relevance judgment and provide some explanations for his or her judgment. At this stage, the query case description and the candidate case document were presented one by one. The participant was instructed to highlight the contents (e.g., individual words, phrases, whole sentences) that he or she paid attention to while making the relevance judgment. Figure 2 provides an illustration.

S5: Semi-structured Explanation. Further, the participant was instructed to explain his or her relevance judgment within the provided framework. As shown in Table 5, six aspects were provided. Among them, "key element" and "key circumstance" are required since they are the basic components in the provided instructions. The optional aspects are generated by referring to other expert opinions in the legal field. To be specific, along with the "critical factor" (i.e., "key element" and "key circumstance"), the "issue" and "application of law" are the other two aspects that are proposed in the guidance document[7] published by **the Supreme People's Court of China** (SPC). Meanwhile,

[7]http://www.court.gov.cn/fabu-xiangqing-243981.html.

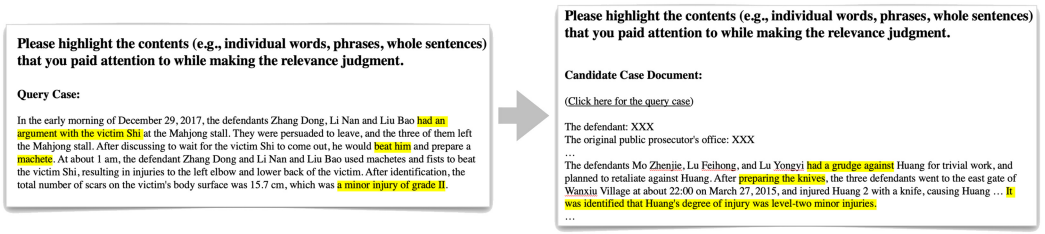


Fig. 2. An illustration of the "Query and Candidate Case Annotation" stage.

comparing the "Cause of Action" is popularly considered as a fundamental step for determining the case relevance in the views of some Chinese legal scholars,[8] and previous research has also pointed out its significant application in the practice of legal case retrieval [44]. The "other" aspect is designed to capture any other potential criteria that have not been mentioned in our study. For each aspect, if the participant considered it when making the relevance judgment, he or she needed to annotate how important this aspect was during his or her decision process via a 5-point scale (1: not important, 5: very important). The participant also needed to give the detailed contents in the query and candidate case regarding this aspect in free texts to support his or her relevance judgment. Otherwise, if the participant did not consider this aspect, he or she could skip the corresponding questions (i.e., the importance score and the detailed reasons). In this stage, the query and candidate cases with the participant's highlights could be referred to anytime. Once completing this stage, the participant could start a new task with the same procedure.

3.4 Experimental Settings

We developed the user study system using Django for participants to log in and complete the experimental procedure. We collected users' relevance judgments (S2 in Section 3.3), explicit feedback (S1 and S3), highlights (S4), and semi-structured reasons (S5) and logged the dwell time of making the relevance judgment via the system. The latest criminal law was provided for reference. The participants could not use other search engines in the study.

Each participant was instructed to complete eight tasks in the study. Specifically, there were 8 experimental conditions in tasks (4 relevance level $\times$ 2 query type), and each participant encountered all the task conditions unconsciously. To ensure that conditions would be completed with equal opportunity, we applied a Latin Square design [17]. In total, each domain expertise group would complete all 64 tasks. Meanwhile, each query-document pair would be annotated by nine participants (three in each domain expertise group). Furthermore, to balance the possible order effects (e.g., learning bias), the tasks were shown in a random order [23]. The participant spent about 1.5 hours completing the main tasks and would gain $18 for involvement.

A pilot study involving three additional users (one of each domain expertise group) was conducted to ensure the study procedure worked well.

3.5 Collected Dataset

After the user study, we noticed that one of the selected query cases remanded,[9] which means that it lacked a final judgment (i.e., golden reference). Therefore, we deleted this query case and the

[8] https://www.chinacourt.org/article/detail/2021/05/id/6050690.shtml.
[9] The query case was selected to involve "Crime of Fraud" and "Crime of Contract Fraud." The case details: https://wenshu.court.gov.cn/website/wenshu/181107ANFZ0BXSK4/index.html?docId=1402cf23d63b4ca3a43baa4900f8e911.

Table 6. Statistics of the Collected Dataset in the User Study

# Query	# Doc (Task)	# Participant	# Session
15	60	72	540

corresponding sessions from our collected data. After cleansing, the collected dataset[10] consists of 540 sessions of relevance judgments corresponding to 15 queries (QSC: 8, QMC: 7). Table 6 shows the basic statistics. The dataset is available now.

As stated in Section 3.4, each document was assessed by three participants in each domain expertise group. We calculated the Fleiss's Kappa κ [15] among the three annotators in each group. The κ is 0.4784/0.2503/0.3626 for the CU/NCU/NLU groups, respectively. The CU group reached a moderate agreement ((0.41, 0.60)), while the NCU and NLU groups reached a fair agreement ((0.21, 0.40)). The differences among different levels of domain expertise will be further discussed in the following sections. Besides, the agreement among the criminal annotators is close to that of LeCaRD (0.500 among three criminal experts [27]), which also suggested that the experimental setting for relevance judgment was appropriate.

4 FACTORS AFFECTING RELEVANCE JUDGMENTS

4.1 Data Analysis Method

Independent Variables. Regarding **RQ1**, we mainly inspect three groups of **independent variables (IVs)**, including *domain expertise*, *query type*, and *case relevance*, from the perspectives of user, task, and result, respectively. As for domain expertise, previous works [31, 57] usually investigated it according to whether the user majors in a specific field (e.g., medicine, finance, politics) and classified users into "in-domain" and "out-domain" groups. Meanwhile, in the legal field, users might still vary in domain expertise depending on their legal specialties [44]. Therefore, we consider the general majors and legal specialties simultaneously in our study. In particular, the NLU group comprises users without law backgrounds, while the CU and NCU groups both major in law. Furthermore, the CU and NCU groups differ in their legal specialties. Since the tasks are criminal cases, we consider the CU group of the highest domain expertise, followed by the NCU and the NLU groups. As for the query type, we divide tasks into the QSC and QMC groups (see Section 3.1) based on the number of involved causes, which is a vital variable in legal practice [44]. As for the case relevance, we use the relevance labels in the dataset as the IV. Following previous research [6, 44], we inspect a binary variable and thus transfer the four-level scores into binary labels based on the relevance instructions in Table 3. In detail, the cases labeled as 1&2 are seen as **Not Relevant (NR)**, and those labeled as 3&4 are **Relevant (R)**.

Measures. We evaluate the performance of users' relevance judgments through accuracy [6] and agreement (Fleiss's Kappa). Specifically, the accuracy metric is calculated by comparing the relevance scores annotated by the participants with the relevance labels. The accuracy metric based on the original four-level scale and the merged binary scale are both examined, denoted as **ACC (4L)** and **ACC (2L)**, respectively. Besides the performance measures, we inspect the process of making relevance judgments through explicit user feedback and implicit user behavior. The explicit user feedback is collected in the pre-task and post-task questionnaires, including self-reported pre-task difficulty, post-task difficulty, and confidence (denoted as **Pre-D**, **Post-D**, and **Conf**). As for user behavior, we mainly look at the **speed** of making relevance judgments, calculated as $\frac{DocLength}{DwellTime}$.

[10]Note that LeCARD updated relevance labels of several cases after we completed the user study. Among the updated cases, five are included in our dataset. The following results we report are consistent with the latest version of LeCARD.

We inspect speed instead of dwell time on the relevance judgment page to avoid the potential bias caused by case length.

Methods. With the observation that all the measures follow a non-normal distribution through the K-S test [25], we mainly use non-parametric statistical tests, except for the Likert scales (i.e., Pre-D, Post-D, and Conf). Regarding the non-parametric statistical tests, we conduct the Mann-Whitney U test [30] instead of the t-test to examine the effects of query type and case relevance, respectively. The Kruskal-Wallis [22] instead of ANOVA is employed to examine the differences among domain expertise groups. Furthermore, the difference between each domain expertise pair is detected by the posthoc Dunn's Test with **Bonferroni-Holm (B-H)** adjustment [13, 20]. Meanwhile, regarding the Likert scales, as previous research [8] has pointed out, it is much more appropriate to summarize them using means and standard deviations, and it is more appropriate to analyze them using the parametric techniques no matter whether following a normal distribution. Therefore, we conduct the corresponding parametric statistical tests (i.e., t-test, ANOVA) on the measures in the Likert scales. All the tests are two-tailed. The p-values are calibrated through the B-H correction [20] within each independent variable to deal with the multiple comparison problem.

4.2 Results

Results are shown in Table 7. The mean value of each measure is reported.

4.2.1 Effects on User Feedback. Among the explicit user feedback measures, the pre-task difficulty is designed for validating the experimental settings. As expected, a significant difference is observed among domain expertise groups. In particular, users without law backgrounds perceive significantly greater difficulty before making relevance judgments. According to the post-task questionnaires, the perceived difficulty decreases a little compared with what they expected in all user groups, while the NLU group still reports significantly greater difficulty and lower confidence than the other two groups. The result indicates that making relevance judgments between legal cases is quite challenging for users without law backgrounds from the user's subjective perspective. However, users with different domain expertise do not show significant differences in relevance judging speed. Concerning the *query type*, it does not have significant effects on any explicit and implicit user feedback measures. Since the causes are not provided to users in query cases, users might not feel the difference explicitly. Regarding *case relevance*, it is reasonable that no significant difference exists in pre-task difficulty since the candidate case has not been shown at this stage, and its relevance should not make a difference. After making relevance judgments, users report significantly less confidence when encountering irrelevant cases. Despite this, the judgment speed just dropped slightly ($p = 0.04$ before B-H correction) in the NR circumstance, but the difference is not distinguishable statistically. In summary, we do not observe any significant differences in the judgment speed under different conditions, such as domain expertise and case relevance, although these factors do cause differences in users' subjective feedback.

4.2.2 Effects on Performance. As shown in Table 7, significant differences can be observed among different *domain expertise* groups in terms of accuracy, including both scales. Remarkably, the users without law backgrounds make many more mistakes than those with legal knowledge. The users majoring in criminal law achieve the highest accuracy among them. The accuracy of the NCU group drops a bit compared with the CU group, but the difference between these two groups is not significant by post hoc Dunn's Test. We explain that the users majoring in law have the fundamental knowledge of the primary legal concepts for determining the case relevance, although the legal specialties would affect their understanding of some detailed points in a specific

Table 7. Effects of Domain Expertise, Query Type, and Case Relevance on the Measures of User Relevance Judgments

Measures	Domain Expertise				Query Type			Case Relevance		
	CU	NCU	NLU	sig.	QSC	QMC	sig.	NR	R	sig.
Kappa	0.4784	0.2503	0.3626	nan.	0.3503	0.3227	nan.	nan.	nan.	nan.
ACC (2L)	0.9189	0.8720	0.7984^{1}	**	0.8958	0.8254	**	0.7893	0.9319	***
ACC (4L)	0.6250	0.5715	$0.4643^{1,2}$	***	0.5938	0.5159	*	0.5057	0.6057	**
Speed	96.32	72.26	88.22	–	98.99	70.30	–	83.96	87.14	–
Pre-D	2.589	2.667	$3.050^{1,2}$	***	2.715	2.829	–	2.793	2.746	–
Post-D	2.572	2.539	$2.911^{1,2}$	***	2.611	2.746	–	2.759	2.595	–
Conf	3.850	3.789	$3.406^{1,2}$	***	3.753	3.600	–	3.598	3.760	*

*/**/*** indicate the difference of domain expertise (query type) is significant at $p < 0.05/0.01/0.001$. The superscripts "1/2" denote that the difference between the CU/NCU group is significant $p < 0.05$ by post-test after B-H adjustment. "nan." indicates the value is unavailable.

case. However, it seems much more difficult for users without law backgrounds to understand the legal relevance in legal case retrieval. The result differs from the study of e-discovery [54], where users with and without law backgrounds did not show significant differences in accuracy metrics. It might be explained by the difference between the two search scenarios. In the e-discovery, the candidate documents are "electronically stored information," including a wide range of document genres, such as letters, memos, emails, and so forth. However, in legal case retrieval, documents are cases decided in law, requiring more professional knowledge to understand and make relevance judgments.

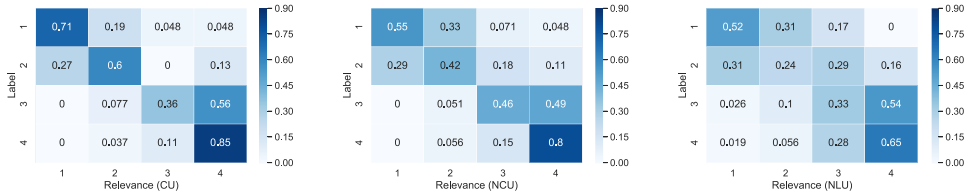
Although no significant differences are observed in user feedback regarding *query type*, it significantly affects the accuracy and agreement among users. As shown in Table 7, the accuracy, especially in the binary scale (2-L), drops significantly in the multiple cause settings. According to the performance measure, the QMC task is more challenging for relevance judgment. As for the effects of *case relevance*, the accuracy metrics in both scales decrease significantly on the condition of irrelevant cases. As a result, the judgment for irrelevant cases might involve more uncertainty.

Regarding the agreement among annotators (measured by Fleiss's Kappa), we observe a dramatic drop in the NCU and NLU groups. Especially for the NCU group, although they make similar performance to the CU group in the accuracy measures, they perform the worst in terms of agreement. We take a detailed investigation into how the annotators disagree with each other. The relevance judgments made by three annotators can be divided into three groups, denoted as "AAA," "AAX," and "AXY," respectively. In detail, the "AAA" denotes that all three annotators give the same relevance score, indicating perfect agreement. "AAX" denotes that only one annotator disagrees with the others, indicating partial agreement, which could be usually solved by the majority vote. "AXY" denotes that the three annotators make totally different relevance judgments, indicating severe disagreement. As shown in Table 8, the CU group can reach at least partial agreement in almost all the cases, and nearly half of the cases achieve perfect agreement. However, when it comes to the NCU group, only a quarter of cases can achieve perfect agreement, while over 10% cases involve severe disagreements. It is not good news for collecting labels since this type of disagreement could not be simply solved, such as by the majority vote. Consequently, additional discussions among the annotators or more annotators might be introduced to solve these disagreements. Interestingly, the NLU group achieves better agreement than the NCU group. We assume that the NLU group might make the same mistakes in some cases, which will be analyzed in the following sections. Furthermore, we investigate how the agreement among each group changes with the query type since we also observe a slight drop in the Fleiss's Kappa measure. As a result,

Table 8. Inter-annotator Agreement in Different Domain Expertise Groups

	# Cases			Ratio in QSC Tasks			Ratio in QMC Tasks		
	AAA	AAX	AXY	AAA	AAX	AXY	AAA	AAX	AXY
CU	28	30	2	0.5000	0.5000	0	0.4286	0.5000	0.0714
NCU	15	37	8	0.2188	0.6562	0.1250	0.2857	0.5714	0.1429
NLU	21	33	6	0.3125	0.5313	0.1562	0.3929	0.5714	0.0357

"AAA" denotes that the three annotators reach perfect agreement. "AAX" denotes that only one annotator disagrees with others. "AXY" denotes that the three annotators disagree with each other.



(a) Criminal (b) Non-criminal (c) Non-legal

Fig. 3. Confusion matrix of the CU/NCU/NLU, respectively. The x-axis means the relevance score annotated by users with the corresponding domain expertise (denoted as "relevance"). The y-axis means the relevance label of each case in the dataset (denoted as "label").

the CU group shows better agreement in the QSC tasks, while more disagreement, especially severe disagreement, occurs in the QMC situation. However, the effects seem indistinguishable and mixed for the NCU and NLU groups.

4.2.3 Error Analysis. Beyond the accuracy values, we conduct a detailed analysis of how the relevance scores assessed by users differ from the labels in the dataset. Figure 3 shows the confusion matrix of each domain expertise group. The darker the diagonal area is, the better the consistency between user assessments and labels. Given the 4-point relevance scale, agreements drop with the decrease of domain expertise in general. Furthermore, if the user with a law background disagrees with the label, the deviation always happens in the adjacency within the same binary relevance interval (i.e., the deviation within 1-2 or 3-4). The deviation is explainable since a case usually contains a variety of detailed circumstances in practice. Users might vary in determining what is significant for relevance judgments. Meanwhile, the deviation is always within the same binary relevance interval, indicating that users with law backgrounds can always distinguish the key elements and accurately make the overall relevance judgment. However, the deviation does not maintain within the same binary interval for the users without law backgrounds. In other words, the mistakes are more severe. For instance, for the cases of somewhat relevant (labeled as 2), the NLU group annotates them as fairly relevant or highly relevant with a non-trivial probability. The results also suggest that users without law backgrounds might lack fundamental ideas of legal relevance and could hardly make relevance judgments accurately for legal case retrieval. Moreover, the mistakes of the NLU group occur more frequently when they deal with irrelevant cases, which suggests that it is more difficult for them to make assessments under this condition.

It is worth mentioning that the NLU group achieves better inter-user agreement than the NCU group, indicated by Fleiss's Kappa in Table 7. We think that users in the NLU group might make the same mistakes in some cases. In order to understand how the mistakes happen, we conduct a detailed case study. In detail, we select the cases that three users of NLU give the same relevance

score but the score differs from the label. Then we examine their reasons manually. As a result, when the deviation occurs in the overall relevance judgment (e.g., 1-3, 2-4), confusing causes are usually involved in the tasks, leading to users' mistakes. For example, they might hardly distinguish the specific causes belonging to the same category. In terms of the confusing causes across categories, they could hardly identify the critical differences between them, and their relevance judgments are mainly based on the matching of keywords. As for the deviation within the same binary relevance interval (e.g., 1-2 or 3-4), it is difficult for them to determine the importance of different circumstances and identify the significant ones related to the constitutive elements of the crime. Besides, they tend to emphasize the sentences that involve specific numeric amounts as the reasons, which are not essential in this task most of the time. In conclusion, they tend to make relevance judgments based on the causes that can be easily identified or keywords matching. In contrast, the instructions based on "key element' and "key circumstance" are not applicable well for the users without law backgrounds.

4.2.4 Discussion. One possible implication of our findings is inspiring the construction of reliable labeled datasets for legal case retrieval. For instance, domain expertise is a critical factor, and thus, the reliable labels should be made by annotators with law backgrounds. Furthermore, the legal specialties also matter. In our study, although the NCU and CU groups do not show significant differences in the accuracy metrics, the NCU group involves rather more disagreements. In that way, if the annotators were not majoring in the corresponding legal specialty, involving more annotators or more discussions would be needed. Besides, the query type has some influence on the quality of relevance judgments, indicating that different strategies might be applied correspondingly. Our results also show that more uncertainty occurs when users encounter irrelevant cases, suggesting that there might be a larger proportion of false-positive annotations when constructing a dataset. On the other hand, it would also be an interesting future direction to collect large-scale labels given these different conditions and re-evaluate the retrieval models for this task.

4.2.5 Summary. Regarding **RQ1**, domain expertise, query type, and case relevance are influential factors for relevance judgment in legal case retrieval. The domain expertise influences subjective user feedback and objective performance. Specifically, it is a much more challenging task for users without law backgrounds. They make significantly less accurate relevance judgments compared with professional users. When dealing with query cases involving multiple causes, the accuracy and inter-user agreement drop, although users do not report significant differences. Last but not least, it seems more challenging to make judgments when encountering a potential irrelevance case, which indicates the corresponding judgment might involve more uncertainty.

5 THE USER VIEW OF RELEVANCE

Regarding **RQ2**, we attempt to investigate users' understanding of relevance according to their semi-structured reasons and text annotations.

5.1 Criteria for Relevance Judgments

To understand the actual users' criteria for relevance, we inspect their semi-structured explanations. We only consider the sessions in which the user gives concrete contents in the corresponding text field for the optional aspects.[11] The external legal expert (Ph.D, majoring in criminal law) annotates the correctness of the written content in each aspect based on the courts' decisions.

[11]Note that the participant skipped the questions if he or vbshe did not consider the corresponding aspect when making the relevance judgment, as we mentioned in Section 3.3.

Table 9. Importance of Different Aspects

	Required		Optional					
	KE	KC	Cause		Issue		AoL	
CU	4.478***	4.006	4.344	[3.994, 4.694]	2.900	[2.372, 3.427]	4.210	[3.713, 4.708]
NCU	4.378***	3.922	3.600	[3.400, 3.800]	3.100	[2.474, 3.726]	3.750	[3.200, 4.300]
NLU	4.378**	4.150	3.421	[3.051, 3.791]	3.800	[3.181, 4.519]	2.000	nan.
ALL	4.111***	4.026	3.776	[3.608, 3.944]	3.160	[2.771, 3.549]	3.909	[3.529, 4.289]

KE/KC/AoL are abbreviations for Key Element, Key Circumstance, and Application of Law, respectively. The average values and the 95% confidence intervals (for optional aspects) are reported. **/*** denotes the difference is significant between "KE" and "KC" at $p < 0.01/0.001$.

Table 10. Usage of Optional Aspects

	Cause		Issue		AoL	
	#correct	#total	#correct	#total	#correct	#total
CU	26	32	10	10	14	19
NCU	55	65	3	10	3	12
NLU	10	19	2	5	0	2

"#total" denotes the number of sessions that the user reports using this aspect, and "#correct" denotes the number of sessions that the user gives correct reasons.

As a result, besides the provided aspects (i.e., key element, key circumstance, cause, issue, and application of law), no other effective aspects for relevance judgment are proposed in the study. Specifically, the reasons in the "other" area are mostly detailed interpretations for the reasons ahead.

The importance of each provided aspect is as shown in Table 9. In general, users of all domain expertise report to follow the instructions according to the importance scores. Recall that the required two aspects (i.e., KE and KC) are used in the relevance instructions. Comparing the two aspects, the "key element" is significantly more important than "key circumstance." The results reflect that users could realize the roles of these two aspects in relevance judgment, where "key element" is more general and qualitative while "key circumstance" is more specific.

Comparing the importance scores of the optional aspects, we note that the overall trend of importance is consistent in the CU and NCU groups, though the differences in the importance scores are more slight among the NCU group. The results indicate that users majoring in law can generally understand the meanings and roles of these aspects in determining case relevance, but users lacking specific criminal knowledge might hardly distinguish among them. However, the trend of importance assigned by NLU groups is contrary. For one thing, they might hardly understand the actual legal meanings of these aspects. For another, this group considers these aspects much less often, as shown in Table 10.

Table 10 provides more details about the usage of the optional aspects, including the frequency and correctness. Among the optional aspects, "cause" is utilized the most frequently. Since the "cause" is the standard expression of criminal charges in the study, which is the legal characterization of the case, it helps determine the relevance between cases. Comparing among the domain expertise groups, users with law backgrounds can identify the causes correctly with a high probability, and the NCU group uses this aspect more often. The results are reasonable. As a fundamental legal concept, the cause is not too difficult for users majoring in law to identify and utilize. Without much more specific knowledge of the criminal law, the relationship of the

Table 11. Overlap between User Highlighted Contents

	Domain Expertise				Query Type			Case Relevance		
	CU	NCU	NLU	sig.	QSC	QMC	sig.	NR	R	sig.
Query	0.7338	0.7233	$0.6876^{1,2}$	***	0.7181	0.6953	***	0.7051	0.7203	–
Candidate	0.6260	0.6022	0.6058	–	0.5831	0.5581	*	0.5479	0.5936	***

The meanings of "*/**/***" and superscripts "1/2" are the same as those in Table 7.

causes works as a significant aspect for determining case relevance among the users majoring in other legal specialties. Meanwhile, users majoring in criminal law can better understand more fine-grained points than the cause (e.g., key circumstances) and thus less refer to the cause. However, identifying the cause is still difficult for users without law backgrounds according to the accuracy and frequency of usage of the cause in the NLU group. As for the "issue," it seems pretty difficult for both NCU and NLU to generalize. Last but not least, although the "application of law" is clearly defined, users still understand it distinctly. This result suggests that "application of law" is still too general to apply in practice.

To sum up, users with law backgrounds can understand legal relevance better and make use of various legal aspects consistently. On the other hand, it is difficult for users without law backgrounds to comprehend the legal meanings of these concepts, and thus their understanding of case relevance might differ from the requirements in the law.

5.2 User Attention

Besides the general relevance criteria, we attempt to understand users' cognitive process of relevance judgment in a fine-grained way. In our user study, participants were instructed to highlight the contents they paid attention to while making relevance judgments. We consider the explicit text annotations by highlighting as an indicator of user attention following previous research [6, 24, 33]. We study the consistency of text annotations under different conditions and then investigate the patterns of user attention distribution based on these annotations. In particular, we inspect how different biases influence the attention allocation during relevance judgment, including the *positional bias*, which is popularly discussed in general web search [24], and the *structural bias*, which is caused by the internal structure of a legal case.

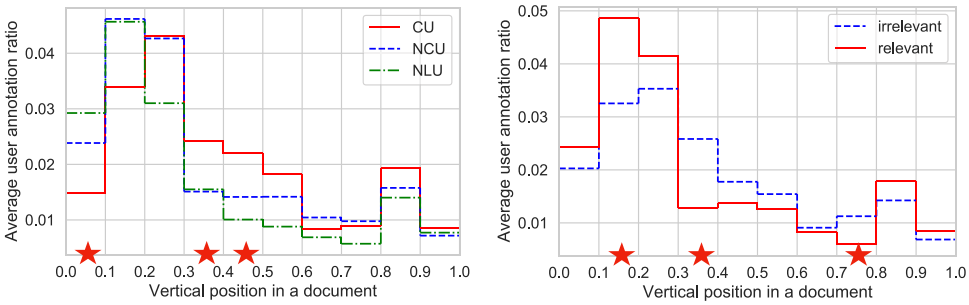
5.2.1 Consistency. To measure the consistency of the text annotations, we employ the overlap coefficient [51], which enables us to compare the annotations of different lengths, following previous studies [6, 33]. The metric is calculated as follows:

$$\text{Overlap}\,(A_1, A_2) = \frac{|A_1 \cap A_2|}{\min\left(|A_1|, |A_2|\right)}, \tag{1}$$

where A_1 and A_2 are two sets of words annotated by two users. In our study, we split the case document into words using the Chinese word segmentation toolkit[12] and filter out the Chinese stopwords. If the user highlighted partial words, we consider the whole words annotated. The overlap coefficient is calculated between each pair of users for each query or candidate case. Table 11 gives the average values of the overlap coefficient. Similar to Section 4, we mainly investigate the effects of domain expertise, query type, and case relevance on this metric.

As shown in Table 11, there exist significant differences among domain expertise groups in the consistency of query case annotations, and the consistency in the NLU group is significantly lower than the other two user groups. As expected, users seem more confused and pay attention to di-

[12]https://pypi.org/project/jieba/.



(a) Domain Expertise (b) User Relevance

Fig. 4. Average annotation ratio in vertical intervals. Differences in the intervals marked by "star" are significant (after B-H correction).

verse contents in the face of the query case description without legal background compared to those majoring in law. However, all domain expertise groups achieve similar intra-group agreements in candidate case highlights. The result indicates that although different domain groups vary in understanding relevance criteria, users show consistent patterns when comparing two cases, such as matching. On the other hand, the number of causes involved in the query case significantly impacts the consistency of user annotations in both query and candidate cases. We assume that a query case might be more complicated if multiple causes are involved, and thus users might focus on different contents in the whole case description. The difference in candidate case annotations is less significant than that in query case annotations. It might be because users would consider the matching signals between two cases more in candidate case annotations, which might weaken the influences of divergence in query case understanding. As expected, the relevance label of the candidate case has no significant effects on the consistency of query case annotations, which also validates our experimental settings. However, there appear more disagreements in user attention when they judge an irrelevant case. It is reasonable that the evidence for determining irrelevant documents might be more dispersed.

5.2.2 Distribution in Vertical Position. Based on the highlights, we further investigate the positional patterns of user attention allocation in the candidate cases. Since users might annotate words or sentences, we consider the short sentence as a unit here to reduce potential biases. In detail, each case is segmented into a list of short sentences by the comma. We use the comma instead of the period because the whole sentence (split by period) that contains multiple circumstances might be pretty long, while a short sentence (split by comma) usually involves a single point. Then, the short sentences that include highlighted texts are denoted as "1" and the other as "0." In that way, we could obtain a vector composed of 0 and 1 for a candidate case based on a user's annotation. In total, we construct 540 vectors for all the candidate cases in our study. To observe the distribution on vertical positions, we divide each vector into 10 bins evenly and consider the proportion of "1" in each bin as the annotation ratio. In our study, each case is divided into 10 vertical intervals, as shown in Figure 4.

Generally, the vertical intervals can be grouped into three areas, i.e., the top 30%, 30%–80%, and the last 20%. The first 30% attract the most attention, and then the ratio drops a lot after 30%. In the intervals from 30% to 80%, the ratio always decreases gradually. Interestingly, the ratio shows a sharp increase at the beginning of the last 20% intervals. We explain these patterns by combining the position bias and the document structures. On the one hand, users tend to read the beginning document for preliminary relevance judgment, and user attention decays with the height, similar

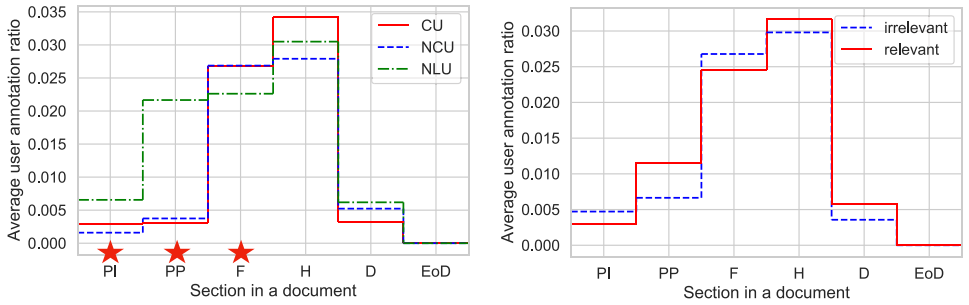
to the previous works in web search [24]. On the other hand, a case document is semi-structured, usually composed of six basic components (also referred to as **sections**) [44], i.e., *Party Information*, *Procedural Posture*, *Facts*, *Holdings*, *Decision*, and *End of Document*. Since the first two sections (i.e., Party Information and Procedural Posture) usually consist of a small proportion of sentences in the whole document, the beginning of the "Facts" will occur in the latter part of the top 30% area. Given that, users might pay much attention to this area for an overview of the case. In particular, the "Facts" contains more detailed information following the case summary, such as arguments from both sides, evidence, and so forth. Compared with the case summary, the detailed information might be less important for users to judge, which also explains the decrease of the annotation ratio in the middle area. It is worth mentioning that there is an increase in the 80%–90%. We think it might be because this area usually involves contents of "Holdings" and "Decisions," which are court opinions and should be a significant reference for relevance judgment.

Furthermore, we investigate the effects of domain expertise on the distribution. Results are shown in Figure 4(a). We conduct the Kruskal-Wallis test in each interval as well as the B-H adjustment for p-value. As a result, we observe significant differences in the 0%–10%, 30%–40%, and 40%–50% intervals after the adjustment. Users with lower domain expertise seem more likely to be affected by position bias. For instance, the NLU group annotates more content at the very beginning of the document, and the corresponding ratio drops earlier and faster. As for the middle area of the case (i.e., 30%–50%), the CU group pays more attention than the other two groups. With more specific knowledge of criminal law, the users might further compare more detailed information beyond the brief summary of the case for making relevance judgments.

In this part, we also wonder whether users' attention allocation will differ when they consider a candidate case to be relevant or not. As shown in Figure 4(b), we observe significant differences in several intervals by Mann-Whitney U Test with B-H adjustment. Generally, the change of the annotation ratio with the vertical positions seems sharper when the case is relevant. In particular, users pay more attention to the top area that usually involves the case summary and less attention to the areas that contain details if they think the candidate is relevant. We think that users might be confident when they judge a candidate to be relevant and thus mainly focused on the general but key points. On the other hand, they might be more cautious and consider more details for their irrelevance judgments.

5.2.3 Distribution in Components of the Case Document. We further investigate the annotation ratio in different components with the assumption that the internal document structure would influence user attention allocation. Similarly, we segment texts in a component (i.e., section) into short sentences by the comma and calculate the annotation ratio, i.e., the proportion of highlighted sentences in each. Results are shown in Figure 5. Generally, the "Facts" and "Holdings" are the principal parts of a case document and tend to draw the most user attention in our study. Specifically, the "Facts" describe the full case story and the "Holdings" contain the court's analysis of the case, which are fundamental to determining case relevance. Compared with the "Facts" section, the "Holdings" section is usually more concise, including key points for the court to make decisions, and thus involves the highest annotation ratio. On the other hand, the "Decision" section that incorporates the final sentence might be too general to compare the relevance between cases, though it is always considered as a core part in a case document.

As shown in Figure 5(a), different distributions occur across domain expertise groups. Users without law backgrounds show different behavioral patterns compared with those majoring in law, especially in the first three sections. They allocate much more attention to the "Party Information" and the "Procedural Posture" than other user groups. These sections are mainly composed of the basic information of both sides and former backgrounds, rarely mentioning the concrete



(a) Domain Expertise (b) User Relevance

Fig. 5. Average annotation ratio in sections of the case document. Differences in the sections marked by "star" are significant (after B-H correction). PI/PP/F/H/D/EoD denotes Party Information, Procedural Posture, Facts, Holdings, Decision, and End of Document, respectively.

story of the current case, and thus seem less significant for relevance judgment. Meanwhile, the NLU group shows less interest in concrete case circumstances than the other groups. We assume that it is difficult for users without professional legal training to identify key points in the lengthy case story description. Given that, they might prefer contents that involve explicit legal items (e.g., charges), even though some are not indeed related to the current case, such as the previous verdict or criminal records mentioned in the "Party Information" and "Procedural Posture" sections. These different patterns also indicate that users without law backgrounds might focus on information that is less helpful for the relevance judgment task. On the other hand, we observe similar patterns in sections under different relevance conditions, shown in Figure 5(b). Combined with the observations in vertical positions (see Figure 4(b)), we think that the general attention allocation on the broad sections is similar, though it might vary in more specific positions when users think the case is relevant or not.

5.2.4 Summary. We focus on investigating how users allocate attention when making relevance judgments based on their highlights, including inter-user consistency and distribution patterns. Regarding consistency, more disagreements occur in the query case understanding when users lack domain knowledge or multiple causes are involved. Meanwhile, an irrelevant case might involve more inconsistent user attention than a relevant one. Regarding the attention distribution, we observe the impacts of both positional and structural biases. In vertical positions, users tend to pay much attention to the top 30% parts, followed by a sharp drop. The middle area (i.e., 30%–40%) is less cared about, and the annotation ratio decreases gradually in this area. There is an increase in the last area of the document, which might be related to the document structure. Users mainly focus on the "Facts" and "Holdings" parts, considering the case structure. Furthermore, these effects on the attention distribution patterns also differ with domain expertise and case relevance. One of the challenges in legal case retrieval is to process the lengthy legal case documents [43, 56], and we believe that these findings can provide inspirations for the related research.

6 THE SYSTEM VIEW OF RELEVANCE

In this section, we focus on system relevance. To address **RQ3**, we first compare the distribution of attention weights in retrieval models with that of users. Then we attempt to improve the performance of retrieval models with the help of user attention.

6.1 Model Attention vs. User Attention

We consider two categories of models, including traditional **bag-of-words** (BOW) models and dense models. Specifically, we inspect the tf-idf [38] and BM25 [36] among the BOW models and BERT [12] among the dense models. These models are representative in each category and popularly adopted in legal case retrieval [27, 34, 37, 43].

6.1.1 Experimental Settings. In addition to the collected dataset in our study, we use LeCaRD [27] for training and validating. In the following experiments, we denote LeCaRD and the dataset collected in the user study as *Dataset-L* and *Dataset-U*, respectively. As for the BOW models (i.e., tf-idf and BM25), the *idf* is calculated based on all of the documents in LeCaRD. Since cases in the user study are mostly generated from LeCaRD, we think it could represent the vocabulary distribution. The parameters in BM25 are set as default values [35]. As for the dense models (i.e., BERT), we utilize the version that is pre-trained on $6.63M$ Chinese criminal case documents [60] (referred to as *Criminal-BERT*). We then fine-tune it with a binary sentence-pair classification task for relevance prediction. These experimental settings are consistent with those in LeCaRD [27]. In particular, we filter out the query cases selected for the user study along with all of their candidate cases from LeCaRD and then divide the left data into training and validation sets randomly by $4 : 1$. In that way, we have 73 queries for training and the remaining 18 for validating. Under each query case, there are 30 candidate cases with four-level relevance labels. We transfer the four-level labels into a binary scale for simplicity. Similar to the above analysis, the cases labeled as 3 and 4 are relevant, and the others are not relevant in the binary scale. For the relevance prediction task, we utilize the three core sections of a case document, i.e., "Facts," "Holdings," and "Decision," and concatenate the query case with each section, respectively. Since the three sections are from the same case, they share the label. For each section type, we fine-tune a BERT correspondingly. Models for these three sections are trained following the same setting. In detail, we truncate the texts symmetrically for each query-section pair input if it exceeds the maximum input length of BERT. Adam optimizer is used, and the learning rate is set as $1e-5$. According to the validation data, all the models could converge after around two epochs. Note that we use the query-section pair instead of the query-document pair as input. Since the case document is always long, especially the "Facts" section, only a part of "Facts" would be considered in the traditional query-document pair under the length limitation of BERT (i.e., 512 tokens). Given that, we utilize three sections separately in the experiments.

The Dataset-U works as the testing set for all methods. Metrics for ranking and classification are utilized for evaluation. Different from Dataset-L, each query case in Dataset-U only involves four candidate cases. In that case, we focus on evaluating the entire ranking list with MAP. Meanwhile, we also utilize the micro-average of precision, recall, and F1 scores as evaluation metrics, following recent benchmarks for legal case retrieval [34]. Since tf-idf and BM25 methods only give ranking scores, they are evaluated with the ranking metric (i.e., MAP). Meanwhile, the BERT models are training for classification. To calculate the ranking metrics, we rank the results according to the predicted probability to be relevant.

The performance in Dataset-U is shown in Table 12. Note that our focus is to inspect the attention weights rather than compare the performance of models. We look at the performance to validate the experimental settings (e.g., model training) before calculating the specific attention weights. As expected, the BERT models outperform the BOW ones on a non-trivial scale. Comparing tf-idf and BM25, BM25 achieves better performance in both ranking metrics. Comparing the BERT models of different sections, metrics for ranking and classification do not show a consistent trend. We think all three sections are informative for relevance prediction but might not play the same roles. We also analyze them separately in the following experiments. To sum up, the

Table 12. Performance of Relevance Prediction on the Dataset-U

Model		MAP	Prec	Recall	F1
BOW	tf-idf	0.6667	–	–	–
	BM25	0.7500	–	–	–
Dense	BERT-F	0.8037	0.6000	0.6774	0.6364
	BERT-H	0.7889	0.7000	0.6774	0.6885
	BERT-D	0.8315	0.6923	0.5806	0.6316

BERT-F/BERT-H/BERT-D denotes the BERT using Facts/Holdings/ Decision sections as input, respectively.

performance of different models shows a similar trend with those in previous studies [27, 43]. Therefore, we think that the experimental settings are reasonable and further analyze their attention weights.

6.1.2 Model Attention. Similar to user attention, we would like to understand what the models focus on when calculating the similarity score. Therefore, we calculate the attention/importance weight of each term as a representation of the model attention. We provide the details of calculating attention weights. The attention mechanism [49] has been well applied in neural models. In particular, BERT is composed of multi-head transformers, which are structured based on self-attention. In self-attention, each word assigns weights to other words, and the corresponding weight could be interpreted as importance or attention. We extract the attention maps from BERT referring to the visualization tool [52] and use the average value across multiple heads. With concatenating the query and section as input, we can calculate the query-to-query, section-to-section, and query-to-section attention maps. Given the input pair $[CLS] < Q > [SEP] < S > [SEP]$, the attention weight on each term of the candidate section s_j is calculated as

$$attn(s_j) = \frac{\sum_{i=1}^{n} \omega_{i,j}}{n} + \frac{\sum_{k=1,k\neq j}^{m} \omega_{k,j}}{m}, \tag{2}$$

where ω denotes the weights in the attention map, and n and m denote the length of query Q and section S in the input, respectively. Following previous work [6], the former part in Equation (2) represents the attention from the query to a term in the section, indicating the matching signal, while the latter part represents the self-attention weight of the section, indicating the importance of the term within the section. For each term in the section, we investigate its role in relevance prediction by summing the two kinds of attention weight. As for the query terms, assuming that users have no idea about the candidate case when they read the query, we focus on the query-to-query attention for representing the process of query case understanding, represented by Equation (3):

$$attn(q_i) = \frac{\sum_{k=1,k\neq i}^{n} \omega_{k,i}}{n}. \tag{3}$$

Regarding each section type, we use the corresponding BERT model that has been fine-tuned on LeCaRD (i.e., BERT-F/BERT-H/BERT-D) to infer the attention weights. Considering the limited input length, we first segment the query and section into text spans with no more than 254 characters when constructing the input pairs. Once getting the attention weights on each span, we concatenate them to obtain the weights of the query or section. In this way, we can make full use of the entire query or section.

On the other hand, attention is not well defined in the traditional BOW models. Nevertheless, we use the weight of each word to represent its importance in the model. To be specific, the $tf \cdot idf$ value is considered to represent the word importance within a text span (i.e., self-attention within

Table 13. Similarity between Model Attention Weights and User Attention in Query/Candidate Case, Measured by Log-likelihood

Model		Candidate Case				Query
		Facts	Holdings	Decision	Merge	
BOW	tf-idf	−0.4569	−0.4989	−0.4743	−0.4200***	−0.9835***
	BM25	−0.2509	−0.3675	−0.1691	−0.2228***	–
	combine	−0.1599	−0.1871	−0.1129	−0.1462***	–
Dense	BERT	−0.1314	−0.1879	−0.1690	**−0.1197**	**−0.5472**

*** indicates the difference in log-likelihood is significantly different from that of BERT at $p < 0.001$.

a query or a candidate section). As for BM25, given the section containing $\{s_1, s_2, \ldots, s_k\}$ words, the contribution of each word s_j in the matching score is measured by

$$\omega(BM25, s_j) = \begin{cases} idf(s_j) \cdot \frac{k+1}{freq(s_j,S)+k\cdot(1-b+b\cdot\frac{|S|}{avgsl})} & s_j \in Q, \\ 0 & else, \end{cases} \tag{4}$$

where $freq(s_j, S)$ denotes the frequency of word s_j in section S, $|S|$ is the length of section in words, and the $avgsl$ is the average length of sections in the collection. The parameters k and b are set as default [27]. Note that the attention weight is calculated in terms (i.e., characters) for BERT and in words for tf-idf and BM25, and we assign each character of the word with the word weight for the BOW models to align the unit. Last but not least, all the weights of each query/section are normalized to the $[0, 1]$ range by *min-max* for comparability across different models.

6.1.3 Comparison between Model and User. We attempt to compare the attention of models and users by inspecting the similarity of their distributions. Similar to Section 5.2, the distribution of user attention is represented by their text annotations. In detail, for each term in the query or section, "1" denotes being highlighted, and "0" denotes the opposite. Taking the "0/1" vector as the representation of the user attention observation, we measure the similarity between two vectors via log-likelihood, inspired by the evaluation of click models. The log-likelihood $ll(m, u, t)$ between the model attention and the user attention on a text span is calculated as follows:

$$ll(m, u, T) = \frac{1}{|T|}\sum_{i=1}^{|T|}(o_{ui}\log\hat{\omega}_{mi} + (1 - o_{ui})\log(1 - \hat{\omega}_{mi})), \tag{5}$$

where o_{ui} denotes whether the user u highlights the $i$the term, $\hat{\omega}_{mi}$ denotes the normalized attention weight of model m, and $|T|$ refers to the length of the text span in terms. To ensure the numerical stability, the model weight $\hat{\omega}$ is clipped between 0.00001 and 0.99999 in Equation (5).

First, we inspect the similarity between model attention and all users' on the query and candidate cases, as shown in Table 13. Besides similarity in each section, we concatenate three sections according to the original order in the case document and measure the overall similarity in the case (referred to as "Merge"). Furthermore, we also average the weights of BM25 and tf-idf in the candidates to consider the internal term importance and matching signal simultaneously (referred to as "combine"). As for the query case, since the query span in the input of the three BERT models is identical, we average their attention weights. Results are shown in Table 13, where the higher value of log-likelihood indicates the higher similarity with user attention. In the query case, BERT outperforms tf-idf significantly in terms of similarity with user attention. It suggests that the dense model (e.g., BERT) might be better in query understanding, while the frequency-based models could hardly identify the essential information in the query case. In the candidate case, we

Table 14. Differences in Similarity between Model Attention Weights and User Attention w.r.t Different Domain Expertise, Query Type, and Prediction Correctness

	Model	Domain Expertise				Query Type			Correctness		
		CU	NCU	NLU	sig.	QSC	QMC	sig.	False	True	sig.
Query	tf-idf	−1.063	−1.017	$-0.8622^{1,2}$	**	−0.9659	−1.0092	–	nan.	nan.	–
	BERT	−0.6004	−0.5747	$-0.4565^{1,2}$	***	−0.5139	−0.5886	**	−0.5497	−0.5462	–
Cand.	combine	−0.1569	−0.1521	$-0.1274^{1,2}$	***	−0.1478	−0.1442	–	nan.	nan.	–
	BERT	−0.1283	−0.1216	$-0.1078^{1,2}$	***	−0.1235	−0.1150	–	−0.1295	−0.1156	**

The meanings of "*/**/***" and superscripts "1/2" are the same as those in Table 7. The attention weights in the candidate case are calculated by merging the three sections in Table 14.

find that BM25 achieves better agreements with user attention than tf-idf, indicating the matching signal should be more vital in determining relevance. Furthermore, combining two models can improve the similarity, which suggests that both word importance and matching signal are useful for relevance judgment. In general, considering all three sections is beneficial except for the "Decision" section results. The exception might be related to the much lower annotation ratio (see Figure 5) and distinct vocabulary from the query case description. Compared with the BOW models, the BERT model still performs better most of the time in terms of consistency with user attention. Specifically, significant tests are conducted between BERT and other models in the "Merge" column of the candidate case, and BERT achieves significantly higher similarity with user attention. Overall, the better agreement with user attention in both query and candidate cases is also consistent with its better performance in relevance prediction in Table 12. It is worth mentioning that the gap between BERT and the BOW model (e.g., the "combine") in the candidate case is not as pronounced as that in the query case. As an explanation, we think the matching signals, which the traditional BOW models can also obtain, perform a dominant role in the candidate case.

Further, we investigate the differences in the attention similarity under different conditions. Similarly, domain expertise and query type are considered as the independent variable of user and task property, respectively. As shown in Table 14, significant differences are observed among domain expertise groups in both query and candidate cases. In particular, both types of models seem to be much more similar to the NLU users in attention distribution. We thus believe that these retrieval models are mainly based on the basic textual features (e.g., keyword matching) and rarely incorporate legal knowledge in relevance prediction, similar to the users without law backgrounds. Unlike domain expertise, the query type factor has few significant effects on the similarity coefficient. The difference only occurs in the BERT model on the query case, indicating that the query case involving multiple causes might be more confusing for models, especially the dense model. Besides, we are also interested in whether there is any difference in model attention when it makes a correct or wrong prediction. As shown in Table 14, the attention distribution of the model[13] is significantly closer to that of the user on the candidate case when it makes a correct relevance prediction.

Given the above observations, we make a further investigation of the attention similarity on the three specific sections of the candidate case. We mainly care about domain expertise and prediction correctness factors since query type seems to have no effects on this similarity coefficient on the candidate case. Results are shown in Table 15. We find that the significant differences of domain expertise or prediction correctness mostly occur in the "Facts" section. As one of the implications, these results inspire us to improve relevance prediction performance by exploiting professional users' attention, especially on the "Facts" section.

[13]Only BERT is inspected here since the BOW models could not output predicted labels here.

Table 15. Differences in Similarity within Each Main Section of the Candidate Case w.r.t Domain Expertise and Prediction Correctness

Section	Domain Expertise				Correctness		
	CU	NCU	NLU	sig.	False	True	sig.
Facts	−0.1396	−0.1344	**−0.1189**^{1,2}	***	−0.1636	**−0.1108**	***
Holdings	−0.1990	−0.1684	−0.1965	–	−0.1689	−0.1967	–
Decision	−0.1715	−0.1692	−0.1658	–	−0.1762	−0.1647	–

The meanings of "*/**/***" and superscripts "1/2" are the same as those in Table 7.

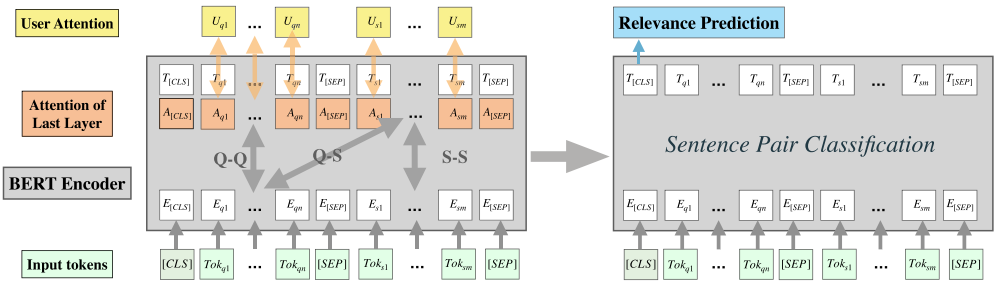


Fig. 6. An illustration of the proposed two-stage framework.

6.2 Proposed Method

Inspired by the above observations, we propose a two-stage framework, which utilizes the attention of users majoring in law for relevance prediction in legal case retrieval. Experimental results on two datasets demonstrate the effectiveness of the proposed methods.

6.2.1 Framework. As illustrated in Figure 6, the proposed framework generally consists of two stages. The first stage aims to optimize the model attention weights with user attention. Given the model that has been tuned in Stage 1, the following stage fine-tunes the model for the target task (i.e., case relevance prediction) with a sentence pair classification task. Details of each stage are given as follows.

In Stage 1, the attention weights are represented in a similar way as described in Section 6.1.2. The attention weights for the terms in a section segment are a combination of query-to-section and section-to-section attention, following Equation (2), denoted as $[A_{s1}, \ldots, A_{sm}]$. Meanwhile, the attentions weights of the query segment (denoted as $[A_{q1,\ldots,A_qn}]$) are based on the query-to-query attention, following Equation (3). On the other hand, we consider the user annotation ratio as the representations of user attention on the query and section segment, denoted as $\boldsymbol{U_q} = [U_{q1}, \ldots, U_{qn}]$ and $\boldsymbol{U_s} = [U_{s1}, \ldots, U_{sm}]$. The annotation ratio of each term is calculated as

$$U_t = \frac{\#\text{Users that highlight the term } t}{\#\text{Users that read this case}}. \tag{6}$$

The object of Stage 1 is to make the model attention close to the user attention, in other words, minimize the loss $\mathcal{L}(\boldsymbol{A_q}, \boldsymbol{A_s}, \boldsymbol{U_q}, \boldsymbol{U_s})$. In particular, we consider three types of loss functions in the following experiments. Taking the raw value of annotation ratio as the observation of user

Table 16. Statistics of Data with User Highlights

Section	Train		Dev	
	# seg	# case	# seg	# case
Facts	635	48	154	12
Holdings	108	34	34	12
Decision	22	11	12	4

attention distribution, we optimize the following L1 loss:

$$\mathcal{L}_{L1}(\boldsymbol{A_q}, \boldsymbol{A_s}, \boldsymbol{U_q}, \boldsymbol{U_s}) = \frac{1}{n}\sum_{i=1}^{n}|A_{qi} - U_{qi}| + \frac{1}{m}\sum_{j=1}^{m}|A_{sj} - U_{sj}|. \tag{7}$$

Furthermore, we apply a softmax function to each attention representation and the post distributions are represented as $\tilde{\boldsymbol{A}}_{\boldsymbol{q}}$, $\tilde{\boldsymbol{A}}_{\boldsymbol{s}}$, $\tilde{\boldsymbol{U}}_{\boldsymbol{q}}$, and $\tilde{\boldsymbol{U}}_{\boldsymbol{s}}$, respectively. Given the normalized distributions, we attempt to minimize the Kullback-Leibler divergence loss (Equation (8)) or the MSE loss (Equation (9)):

$$\mathcal{L}_{KLD}(\boldsymbol{A_q}, \boldsymbol{A_s}, \boldsymbol{U_q}, \boldsymbol{U_s}) = \frac{1}{n}\sum_{i=1}^{n}\tilde{U}_{qi}(\log\tilde{U}_{qi} - \log\tilde{A}_{qi}) + \frac{1}{m}\sum_{j=1}^{m}\tilde{U}_{sj}(\log\tilde{U}_{sj} - \log\tilde{A}_{sj}) \tag{8}$$

$$\mathcal{L}_{MSE}(\boldsymbol{A_q}, \boldsymbol{A_s}, \boldsymbol{U_q}, \boldsymbol{U_s}) = \frac{1}{n}\sum_{i=1}^{n}|\tilde{A}_{qi} - \tilde{U}_{qi}|^2 + \frac{1}{m}\sum_{j=1}^{m}|\tilde{A}_{sj} - \tilde{U}_{sj}|^2. \tag{9}$$

Given the model optimized in Stage 1, we further fine-tune it for relevance prediction with a sentence pair classification task in Stage 2. Following the classic flow, the query-section pair is separated by the [SEP] token to construct the input in the form of $[CLS] < Q > [SEP] < S > [SEP]$. Then the final hidden state vector of the [CLS] token is fed into a fully connected layer to make binary classification, which optimizes a cross-entropy loss. The text pair is truncated symmetrically in this stage if exceeding the maximum length, which makes the result comparable to that in Table 12. Since we focus on investigating the attention mechanism in this article, the more complicated models are beyond the research scope and left for future work.

Different from the process of extracting attention of the fine-tuned model in Section 6.1, the proposed framework could be viewed as a reverse process. It first tunes the parameters by optimizing the attention distributions and then fine-tunes the model for relevance prediction.

6.2.2 Experimental Settings. In the proposed framework, the first stage requires users' highlights as labels, and thus only the Dataset-U is involved. The query cases are divided randomly at 4:1 as training (12 queries and 48 candidates) and validating sets. According to the former analysis of domain expertise, the users without legal knowledge are much more likely to make incorrect relevance judgments. Meanwhile, their attention distribution is also significantly different from those majoring in law. Therefore, we only use the annotations of the CU and NCU groups to construct labels to avoid noisy guidance. In order to make full use of user annotations, we divide the query and candidate section into segments with no more than 254 characters and construct the input based on each pair of query and section segments. In particular, we filter out the input pairs that involve the segment without any user annotation to ensure numerical stability. Table 16 shows the statistics of the filtered dataset used in Stage 1. In this stage, we utilize the Criminal-BERT [60] as the base model. As for the training process, we use the Adam optimizer and set the learning rate as $1e-5$. We select the stopping point according to the loss on the validation set and adopt the

Table 17. Performance of Relevance Prediction on Dataset-U and Dataset-L

Section	Dataset	Model	MAP	Prec	Recall	F1
Facts	Dataset-U	base	0.8037	0.6000	0.6774	0.6364
		ts-L1	**0.8426**	**0.7000**	0.6774	**0.6885**
		ts-KLD	0.8093	0.6111	0.7097	0.6567
		ts-MSE	0.8148	0.6053	**0.7419**	0.6667
	Dataset-L	base	0.7556	0.7299	0.8370	0.7797
		ts-L1	0.7563	0.7392	0.8397	0.7863
		ts-KLD	**0.7589**	**0.7476**	0.8614	0.8005
		ts-MSE	0.7559	0.7426	**0.8859**	**0.8079**
Holdings	Dataset-U	base	0.7889	0.7000	0.6774	0.6885
		ts-L1	0.8056	0.6970	**0.7419**	**0.7188**
		ts-KLD	0.8000	**0.7692**	0.6452	0.7018
		ts-MSE	**0.8389**	0.6571	**0.7419**	0.6970
	Dataset-L	base	0.8685	0.7572	0.9043	0.8242
		ts-L1	**0.8853**	**0.7767**	0.8697	0.8206
		ts-KLD	0.8564	0.7586	0.8777	0.8138
		ts-MSE	0.8624	0.7610	**0.9229**	**0.8341**
Decision	Dataset-U	base	**0.8315**	0.6923	0.5806	0.6316
		ts-L1	0.8037	**0.7917**	**0.6129**	**0.6909**
		ts-KLD	0.7926	0.6471	0.3548	0.4583
		ts-MSE	0.7704	0.7368	0.4516	0.5600
	Dataset-L	base	0.7964	0.8000	**0.7171**	**0.7563**
		ts-L1	**0.8003**	**0.8168**	0.5994	0.6914
		ts-KLD	0.7719	0.8157	0.5826	0.6797
		ts-MSE	0.7862	0.8149	0.6415	0.7179

corresponding checkpoints for Stage 2. The second stage requires the final relevance label for fine-tuning. Therefore, we could also utilize the Dataset-L in this stage. The pre-processing of dataset is the same with as in Section 6.1.1, including data filtering, train/dev sets partition, text-pair truncation, and so forth. Similarly, the Adam optimizer is utilized, and the learning rate is set as $1e-5$ during training. The main difference lies in that the fine-tuning process is conducted on the model saved in Stage 1 rather than the initial Criminal-BERT. According to the loss on the validation set, this stage could always converge after about two epochs, and we adopt the best epoch on the validation set for evaluation. To validate the effectiveness of the proposed framework, we consider the model fine-tuned directly based on the Criminal-BERT as the baseline.

The Dataset-U is considered as the testing dataset for evaluating relevance prediction. Besides, we also compare the performance on the validation set of Dataset-L. Similar to the previous sections, the evaluation metrics include the ranking metric (e.g., MAP) and classification metrics (e.g., micro-precision, recall, F1). The models of each section category are trained and evaluated separately.

6.2.3 Results. The performance of relevance prediction on both datasets is shown in Table 17. Among the models, "base" refers to the baseline model that is fine-tuned directly based on the Criminal-BERT, while "ts" refers to the proposed two-stage method. Specifically, "L1/KLD/MSE" refer to the three types of loss functions in Stage 1, respectively. In general, the two-stage models outperform the baselines based on the "Facts" and "Holdings" sections, suggesting the effectiveness of optimizing model attention via user attention (Stage 1). Moreover, the proposed framework

Table 18. Similarity between Model Attention Weights and User Attention in Query/Candidate Case in Dataset-U

Model	Facts		Holdings	
	Query	Candidate	Query	Candidate
base	−0.5821	−0.3339	−0.5741	−0.2582
ts-L1	−0.5328	−0.3126	−0.5307	−0.2286
ts-KLD	−0.5513	−0.3142	−0.5466	−0.2458
ts-MSE	−0.5513	−0.3185	−0.5573	−0.2442

achieves performance on both datasets, i.e., with user highlights and without. The result is encouraging. Since it is much more time-consuming for annotators to provide fine-grained text annotations than the mere relevance label, the affordable dataset that involves user highlights might be relatively small-scaled. The proposed framework can adapt to the limited data size, where the second stage can utilize more data without user highlights. As a result, it also works well on the data without user highlighting. The result shows that we can exploit the limited user highlights to improve the general legal case retrieval task.

Among different sections, the improvements on the "Facts" section are more outstanding. The result is consistent with the former analysis in Table 15, where the significant differences mainly occur in the "Facts" section. The results on the "Decision" section seem a bit strange. Given that only a tiny proportion of "Decision" sections contain user annotations (see Table 16), we think that the few data are likely to cause misleading (e.g., over-fitting) in the first stage and further hurt the performance of the entire model. Therefore, we mainly look at the "Facts" and "Holdings" sections in the following analysis.

Furthermore, we inspect the attention weights of different models by calculating the similarity with the annotation ratio of the users with law backgrounds (i.e., CU and NCU groups). The similarity is measured via log-likelihood, as described in Section 6.1. Since all the models are trained and evaluated on the truncated texts, we calculate the similarity based on the same texts. Results are shown in Table 18. Compared with the "base" model, the attention weights in our proposed methods are more similar to those of the users. This trend is consistent in both query and candidate cases. The results suggest the effectiveness of integrating user attention into model attention in the proposed method.

6.3 Summary

Regarding **RQ3**, we investigate the similarity between model and user in their attention distribution. Generally, the BERT model is more likely to agree with users in attention allocation than the traditional BOW models on both query and candidate cases. Specifically, we find that the model's attention is more similar to that of users without law backgrounds than that of professional users. Meanwhile, the model attention is closer to the users' when it makes a correct prediction. Inspired by these findings, we propose a two-stage framework that utilizes professional users' attention distribution for legal case retrieval. Experimental results on distinct datasets demonstrate its encouraging improvements.

7 CONCLUSIONS

In this article, we work on understanding relevance judgments in legal case retrieval from multiple perspectives. We conduct a laboratory user study centered on legal relevance that involves 72 participants with distinct domain expertise. With the collected data, we make an in-depth investigation into the process of making relevance judgments and attempt to interpret the user

relevance and system relevance in this search scenario. In particular, we have made several interesting findings with regard to the research questions.

Regarding **RQ1**, we inspect whether the properties of user, query, and result would affect the process of making relevance judgments. In conclusion, the user's domain expertise significantly affects measures of subjective user feedback and objective performance. Specifically, users without law backgrounds are more likely to make mistakes and tend to perceive greater task difficulty. The query type (i.e., the number of causes involved in the query case) seems not to make any difference in user feedback, while the performance drops under the multi-cause condition. As for the result property, we find that users might make more mistakes and feel less confident when they encounter a potential irrelevant case. Besides, it is worth mentioning that the accuracy and inter-user agreement are distinct measures for performance in legal case retrieval. In our study, although the users majoring in law achieve close accuracy measures, the users out of the corresponding legal specialty show greater disagreements in relevance judgments. Meanwhile, we find that the users without law backgrounds might make identical mistakes and thus significantly hurt the accuracy of relevance judgments, though they show better inter-user agreement than some professional users.

Regarding **RQ2**, we investigate how users understand legal relevance based on their semi-structured reasons (Section 5.1) and fine-grained text annotations (Section 5.2). As for the generalized relevance criteria, users report to follow the relevance instructions well and distinguish the importance of "key element" and "key circumstance" in determining case relevance. Besides, the "cause" is sometimes considered to support the decisions, especially by the users specialized in other laws, while "issue" and "application of law" seem less helpful in legal practice. Besides, we observe that users without law backgrounds can hardly understand these legal aspects or their roles in relevance judgments. On the other hand, taking user highlights as the indicator of their attention, we inspect the inter-user consistency and observe various patterns of attention distribution. According to the *Overlap* metric, users majoring in law achieve higher consistency in query understanding, and the multiple causes in the query or the potential irrelevant candidate might involve more disagreements. Different from the general web search [24, 58], the attention distribution in vertical positions can be divided into three groups (i.e., 0%–30%, 30%–80%, and 80%–100%), which might result from positional and structural biases. Furthermore, different patterns of attention distribution can be observed under different domain expertise and relevance judgments.

Regarding **RQ3**, we extract the attention weights of retrieval models and compare them with users' attention. Generally speaking, the neural retrieval model (i.e., BERT) seems to be closer to users than the BOW models in terms of attention distribution. Specifically, the model attention is more similar to users without law backgrounds, who are more likely to make mistakes in relevance judgments. Besides, the similarity between the attention distributions decreases when the model makes incorrect relevance predictions. Last but not least, we propose a two-stage framework that utilizes the attention of professional users for legal case retrieval. The experimental results show its improvements.

Our work sheds light on relevance in legal case retrieval. It has promising implications for related research, such as the construction of datasets and the design of retrieval models. For instance, aware of the effects of domain expertise, relevance annotations for legal case retrieval should be made by users with professional legal training. More discussions or annotators might be included if the annotators are not majoring in the specific area of law. Besides, since the query type and case relevance are also influential factors, they should be considered when collecting labels, such as designing a quality assurance mechanism. Moreover, understanding how users make relevance judgments in the entire case document and their differences from models could further support the development of retrieval models, such as considering the positional and structural biases.

Specifically, the internal structure of the case document also exists in other legal systems [21, 41], where our findings might be exploited. Beyond the legal case retrieval, our methodologies and findings could also benefit other similar professional search scenarios, such as patent retrieval, medical search, scientific literature search, and so forth.

8 LIMITATIONS AND FUTURE WORK

We acknowledge some potential limitations of our work. One limitation lies in the base dataset, LeCaRD [27], based on which the tasks in our study are designed. The dataset is built for legal case retrieval tasks in the Chinese law system, and some results might be retrained depending on different legal systems (e.g., the common law). Meanwhile, the LeCaRD is not perfect, such as containing the query case lacking a final judgment and involving some wrong labels. Since prior cases are not directly cited in the Chinese law system and no case citation information could be utilized, the relevance label in LeCaRD is determined by expert judgments with final court decisions as golden references. Given that, a case is unsuitable to be included in the public labeled dataset if it lacks a final decision (e.g., being remanded). The LeCaRD has also been updated several times to correct some mistakes in its previous versions. In this article, we keep the reported results consistent with the latest version of LeCaRD, even though the update has not influenced the main conclusions.

Another potential limitation lies in that the size of collected data is limited, as in most user studies [31, 44]. Besides, as an attempt to understand the relationship between user relevance and system relevance, the retrieval models considered in this article are mostly fundamental. More complicated retrieval models are beyond the research scope and left for future research. As the approximation of user attention, the highlights might still vary from the real attention.

There are several promising directions for future work. Besides a laboratory user study, a large-scale crowd-sourcing study is promising. In particular, with a larger-scale dataset, it would be influential to further investigate how the factors that affect the dataset construction would affect downstream applications, such as retrieval system evaluation. This article focuses on the general distribution of attention, while it would also be an interesting direction to perform a linguistic analysis to characterize the differences in the "important" content identified by the retrieval models and users. To obtain more precise and fine-grained user attention, an eye-tracking study is also promising. Last but not least, it is still worth further investigating to incorporate the relevance-judging process and domain knowledge into the more complicated retrieval models for this task.

REFERENCES

[1] Marco Allegretti, Yashar Moshfeghi, Maria Hadjigeorgieva, Frank E. Pollick, Joemon M. Jose, and Gabriella Pasi. 2015. When relevance judgement is happening? An EEG-based study. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'15)*. 719–722.

[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] Carol L. Barry. 1994. User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science* 45, 3 (1994), 149–159.

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

[5] Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019. FIRE 2019 AILA Track: Artificial Intelligence for Legal Assistance. In *Proceedings of the 11th Forum for Information Retrieval Evaluation (FIRE'19)*. 4–6.

[6] Valeria Bolotova, Vladislav Blinov, Yukun Zheng, W. Bruce Croft, Falk Scholer, and Mark Sanderson. 2020. Do people and neural nets pay attention to the same words: Studying eye-tracking data for non-factoid QA evaluation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM'20)*. 85–94.

[7] Pia Borlund. 2003. The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research* 8, 3 (2003).

[8] James Carifio and Rocco Perla. 2008. Resolving the 50-year debate around using and misusing Likert scales. *Medical Education* 42, 12 (2008), 1150–1152.

[9] Heting Chu. 2011. Factors affecting relevance judgment: A report from TREC legal track. *Journal of Documentation* 67, 2 (2011), 264–278.

[10] Cyril W. Cleverdon, Jack Mills, and Michael Keen. 1966. Aslib–Cranfield research project. *Factors Determining the Performance of Indexing Systems*, Volume 1, Design, Part 1, Text. Technical Report.

[11] Erica Cosijn and Peter Ingwersen. 2000. Dimensions of relevance. *Information Processing & Management* 36, 4 (2000), 533–550.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Alexis Dinno. 2015. Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. *The Stata Journal* 15, 1 (2015), 292–300.

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

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

[16] Nigel Ford. 2015. *Introduction to Information Behaviour*. Facet Publishing.

[17] David A. Grant. 1948. The Latin square principle in the design and analysis of psychological experiments. *Psychological Bulletin* 45, 5 (1948), 427.

[18] Jacek Gwizdka. 2014. Characterizing relevance with eye-tracking measures. In *Proceedings of the 5th Information Interaction in Context Symposium (IliX'14)*. 58–67.

[19] 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.

[20] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.

[21] Arnav Kapoor, Mudit Dhawan, Anmol Goel, T. H. Arjun, Akshala Bhatnagar, Vibhu Agrawal, Amul Agrawal, Arnab Bhattacharya, Ponnurangam Kumaraguru, and Ashutosh Modi. 2022. HLDC: Hindi legal documents corpus. *arXiv preprint arXiv:2204.00806* (2022).

[22] 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.

[23] 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 (SIGIR'14)*. 113–122.

[24] Xiangsheng Li, Yiqun Liu, Jiaxin Mao, Zexue He, Min Zhang, and Shaoping Ma. 2018. Understanding reading attention distribution during relevance judgement. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM'18)*. 733–742.

[25] Hubert W. Lilliefors. 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* 62, 318 (1967), 399–402.

[26] 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 (SIGIR'21)*. 1622–1626.

[27] 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 (SIGIR'21)*. (Virtual Event), 2342–2348.

[28] Kelly L. Maglaughlin and Diane H. Sonnenwald. 2002. User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. *Journal of the American Society for Information Science and Technology* 53, 5 (2002), 327–342.

[29] 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* 29, 1 (2021), 1–35.

[30] Henry B. Mann and Donald R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* 18, 1 (1947), 50–60.

[31] Jiaxin Mao, Yiqun Liu, Noriko Kando, Min Zhang, and Shaoping Ma. 2018. How does domain expertise affect users' search interaction and outcome in exploratory search? *ACM Transactions on Information Systems (TOIS)* 36, 4 (2018), 1–30.

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

[33] Chen Qu, Liu Yang, W. Bruce Croft, Falk Scholer, and Yongfeng Zhang. 2019. Answer interaction in non-factoid question answering systems. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR'19)*. 249–253.

[34] 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 (JSAI-isAI'19)*. Springer, 34–49.

[35] Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (LREC'10)*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[36] Stephen E. Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*. Springer, 232–241.

[37] Juline Rossi and Evangelos Kanoulas. 2019. Legal information retrieval with generalized language models. In *Proceedings of the 6th Competition on Legal Information Extraction/Entailment (COLIEE'19)*.

[38] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24, 5 (1988), 513–523.

[39] Tefko Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology* 58, 13 (2007), 1915–1933.

[40] Tefko Saracevic. 2016. The notion of relevance in information science: Everybody knows what relevance is. But, what is it really? *Synthesis Lectures on Information Concepts, Retrieval, and Services* 8, 3 (2016), i–109.

[41] Jaromir Savelka, Hannes Westermann, Karim Benyekhlef, Charlotte S. Alexander, Jayla C. Grant, David Restrepo Amariles, Rajaa El Hamdani, Sébastien Meeùs, Aurore Troussel, Michał Araszkiewicz, et al. 2021. Lex Rosetta: Transfer of predictive models across languages, jurisdictions, and legal domains. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law (ICAIL'21)*. 129–138.

[42] Linda Schamber. 1991. Users' criteria for evaluation in a multimedia environment. *Proceedings of the ASIS Annual Meeting* 28 (1991), 126–33.

[43] 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 *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI'20)*. 3501–3507.

[44] Yunqiu Shao, Yueyue Wu, Yiqun Liu, Jiaxin Mao, Min Zhang, and Shaoping Ma. 2021. Investigating user behavior in legal case retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'21)*. (Virtual Event), 962–972.

[45] Stuart A. Sutton. 1994. The role of attorney mental models of law in case relevance determinations: An exploratory analysis. *Journal of the American Society for Information Science* 45, 3 (1994), 186–200.

[46] Arthur Taylor. 2012. A study of the information search behaviour of the millennial generation. *Information Research: An International Electronic Journal* 17, 1 (2012), n1.

[47] 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 17th International Conference on Artificial Intelligence and Law (ICAIL'19)*. 275–282.

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

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS'17)*. 5998–6008.

[50] Brian C. Vickery. 1959. The structure of information retrieval systems. In *Proceedings of the International Conference on Scientific Information (ICSI'59)*, Vol. 2. 1275–1290.

[51] M. K. Vijaymeena and K. Kavitha. 2016. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal* 3, 2 (2016), 19–28.

[52] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418* (2019).

[53] Jianqiang Wang. 2011. Accuracy, agreement, speed, and perceived difficulty of users' relevance judgments for e-discovery. In *Proceedings of SIGIR Information Retrieval for E-discovery Workshop (SIRE'11)*, Vol. 1. Citeseer.

[54] Jianqiang Wang and Dagobert Soergel. 2010. A user study of relevance judgments for E-Discovery. In *Proceedings of the American Society for Information Science and Technology* 47, 1 (2010), 1–10.

[55] Robert Warren. 2011. University of waterloo at TREC 2011: A social networking approach to the legal learning track. In *Text Retrieval Conference (TREC'11)*.

[56] 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 (ISAI-isAI'20)*. Springer, 269–285.

[57] Ryen W. White, Susan T. Dumais, and Jaime Teevan. 2009. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM'09)*. 132–141.

[58] Zhijing Wu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. Investigating reading behavior in fine-grained relevance judgment. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'20)*. 1889–1892.

[59] Eugene Yang, David Grossman, Ophir Frieder, and Roman Yurchak. 2017. Effectiveness results for popular e-discovery algorithms. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law (ICAIL'17)*. 261–264.

[60] Haoxi Zhong, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. 2019. *Open Chinese Language Pre-trained Model Zoo*. Technical Report. https://github.com/thunlp/openclap.

Received 12 October 2021; revised 19 September 2022; accepted 18 October 2022