ABSTRACT

Given a specific query case, legal case retrieval systems aim to retrieve a set of case documents relevant to the case at hand. Previous studies on user behavior analysis have shown that information retrieval (IR) systems can significantly influence users’ decisions by presenting results in varying orders and formats. However, whether such influence exists in legal case retrieval remains largely unknown. This study presents the first investigation into the influence of legal case retrieval systems on the decision-making process of legal users. We conducted an online user study involving more than ninety participants, and our findings suggest that the result distribution of legal case retrieval systems indeed affects users’ judgements on the sentences in cases. Notably, when users are presented with biased results that involve harsher sentences, they tend to impose harsher sentences on the current case as well. This research highlights the importance of optimizing the unbiasedness of legal case retrieval systems.

CCS CONCEPTS

• Information systems → Information retrieval: Presentation of retrieval results.

KEYWORDS

Legal Case Retrieval, Judicial Judgement, User Study

1 INTRODUCTION

Given a specific query case, legal case retrieval systems aim to retrieve a set of case documents relevant to the case at hand. By providing access to a vast collection of past judgments and their associated legal reasoning, they contribute to the development and application of consistent legal principles, ensuring fairness, predictability, and uniformity in legal decision-making. Therefore, legal case retrieval systems have long served as an essential part of both the common law and civil law systems. For example, in China (a country with civil law systems), the Supreme People’s Court explicitly requires all judges to submit relevant cases retrieved by legal case retrieval systems to the court before making the final decision on a specific case.

Previous studies on user behavior analysis have shown that information retrieval (IR) systems can significantly influence users’ decisions by presenting results in varying orders and formats. Because search engines and recommendation systems have served as the major entries, if not the only entries, for people to access information on the internet, how these systems select and organize retrieved results could directly affect what and how information is distributed to downstream users. User’s opinions and decisions are thus likely to be influenced by the IR system’s behaviors and strategies. For example, previous studies have shown that people’s views on certain topics could be significantly different when we present different results on the top of search engine result pages (SERPs) [10].

To this end, we present one of the first studies on how user decisions are influenced by the retrieved results of legal case retrieval systems. Our goal is to provide a quantitative analysis on how the distribution of sentences in the retrieved cases would affect users’ decisions on the sentences of the query cases. Specifically, we conduct experiments on a Chinese legal case retrieval dataset, i.e., LeCaRDv2, and recruited 94 participants who major in criminal
law (82 of which have already passed National Judicial Exam of China) to simulate judges in courts who need to make decisions on query cases based on the fact descriptions and cases retrieved by legal case retrieval systems. We carefully crafted the retrieved case lists for each participant to study whether the changes of retrieval results would affect their final decisions on the query cases. Experimental results show that participants tend to impose more penalties to the accused when they are presented with cases with harsher sentences, despite of the relevance of the presented cases. This highlights the importance of unbiasedness in the optimization of legal case retrieval systems.

Since the fact that legal case retrieval systems have become increasingly prevalent in the past years, studying whether legal case retrieval systems have an influence on users in the field of justice and evaluating this influence can help eliminate biases in legal case retrieval systems and optimize fairness and justice in the field of justice, with far-reaching practical significance and application value.

The remainder of the paper is organized as follows. Section 2 reviews the related works on the influence of information retrieval systems, the bias in information retrieval systems and legal case retrieval systems. Section 3 presents details of the user study we conduct. Section 4 describes the results of the user study. The last section concludes the paper.

## 2 RELATED WORK

Here are some studies on related topics such as the influence of information retrieval systems on their users, the bias in these systems, and the current state of legal case retrieval systems.

### 2.1 IR system’s influence on user behaviors.

Information retrieval systems, including search engine systems and recommendation systems, have a certain influence on users. For example, previous studies show the the use of search engine may make people over-confident [26] and recommendation systems in online market may induce user to purchase more often [4, 21]. Studies on this aspect are abundant.

Some companies have borrowed the idea that recommendation systems have an influence on users’ decision-making process and achieved certain results, for example, previous study shows that simply changing the video ranking to place it before popular videos can make it more popular [5]. Search engine systems also have an influence on users’ decision-making process and cognition. For example, it is found that when the information retrieved by search engine is similar to users’ cognition, they are likely to make the same decision [3, 27]. In the job market, a previous study shows that relevant searches for jobs affect users’ decision-making when looking for jobs [1]. In the study of search engine result pages, it is pointed out that just viewing the top entries returned by the search engine can change users’ views on the search topic [13]. Similarly, it is shown that it is feasible to manipulate search results to change voters’ preferences in an undecided democratic election [8]. Furthermore, there is a study show that even a brief selective exposure to online search results can affect users’ attitudes towards elections [11].

### 2.2 The Bias in IR Systems

There is various bias types in information retrieval systems [2] and there are previous studies focus on the bias itself and its influence attempt to address it and reduce it [12, 17, 22, 28]. A previous study also points out biases present in search engine result pages [20]. Additionally, previous studies in the medical diagnosis field show that this bias has both positive and negative effects [23]. On the user side, biases may arise due to the biases of information retrieval systems. There is a study find cognitive biases in users when using information retrieval systems by asking them to answer questions before and after searching for information and labeling the usefulness of the searched information [14]. Previous studies find that users are very uncertain when evaluating the credibility of search engine results pages and often have different impressions from fact-checking [10]. Previous work demonstrated that search engine result pages can cause users to have biases [20], and it is pointed out that this bias can also affect users’ information retrieval behavior in return [6]. If considering manipulating information retrieval systems, according to the research, although such behavior can influence users’ behavior, users can detect that the search results have been manipulated [7].

### 2.3 Legal Case Retrieval Systems

The work published in the top-level comparative law journal in the United States, *American Journal of Comparative Law*, indicates that although China, as a non-precedent jurisdiction, explicitly prohibits judges from referencing prior cases as the basis for their judgements, and in previous investigations, most judges have made clear that prior cases play only a marginal role in their decisions, some Chinese judges still rely heavily on cases retrieved when making decisions [16]. Since the fact that some judges in China, a statutory law country, rely on the cases retrieved, a credible and well-performing case retrieval system is essential.

There are some previous studies on legal case retrieval systems. The topics are various. One of these studies constructs a large legal case retrieval dataset [19]. While previous studies investigate the user behavior in legal case retrieval to benefit the design of the corresponding retrieval systems to support legal practitioners [25], there are also some studies focus on the ranking to improve the performance of legal case retrieval systems [18, 24].

## 3 USER STUDY

Initially, we clarify our research question. After that, we introduce our experiment platform and online user study, including how we recruit participants, design task data, and collect experiment results.

### 3.1 Research Questions

The research questions we want to address in this paper are:

- Would user’s judgments be affected by the distribution of cases retrieved by the legal case retrieval systems?
- If so, are there any patterns between the distributions of the retrieval results and the user’s sentences to the query cases?

These two questions are the key to understand whether legal case retrieval systems have an influence on their users. If there are special patterns between case retrieval results and user’s judgments
on query cases, we could use them to identify the potential bias in existing retrieval systems and improve the overall quality and fairness of the legal judgment process.

3.2 Methods

The workflow of our experiment is shown in Figure 1. The experiment consists of four tasks and one user questionnaire, with each task requiring participants provide sentences to the query case based on the information (i.e., law descriptions and cases retrieved by the retrieval systems) provided by the platform and their own thinking and cognition. The questions in the questionnaire are listed in Table 1, and the possible answers to each question is completely disagree, disagree, neutral, agree, and completely agree. After completing all tasks, participants are required to complete this questionnaire based on their experiment experience. The design of the experiment tasks simulates the process of judicial judgement in a real judicial judgement scenario. In practice, after conviction, a judge would make a sentence based on the following information:

- Related laws and regulations, including criminal law and sentencing guidelines.
- ‘Extraneous factors’ beyond the case circumstances, including the attitude of the suspect towards confession, the economic situation of the victim, public opinion, and social stability factors.
- Sentences of relevant cases retrieved by the legal case retrieval systems.

Note that a standard judicial judgment process usually consists of two separate parts, i.e., conviction and sentence. For simplicity, we only focus on the sentence part and reveal the actual crime of the legal judgment process.

### 3.3 Dataset

The dataset used in this experiment is LeCaRDv2[15]. The dataset includes several criminal verdicts in China and is organized by query. For each query, the top 30 documents of BM25 is retrieved from a corpus with over 55,000 documents. Overall, the dataset has 800 queries and the average BM25 total score of these top 30 documents is 20.89.

For simplicity, we picked 4 representative queries from the 800 queries for our experiments, and the process of query selection is described as the followings. To begin with, we choose specific 4 crimes, that is, injury, robbery, rape and murder, which are 4 violent serious crimes. For these crimes, at the sentencing level, the range of sentence that are applicable under the criminal code is broad, which is suitable for observing the influence of a case retrieval system on judicial judgement made by users. Additionally, we select only one query for each of these crimes, in order to prevent participants from being influenced by each other if they sentence two different cases with the same crime later on, thereby making it inconvenient to group tasks at the level of trials.

To pick these 4 queries out, we filter the queries that

1. are not in those crimes
2. are longer than 5,000 words
3. are sentenced to probation or death
4. their documents candidates that with the same crime and not sentenced to probation or death are more than 10

Then, for each crime, if its queries are more than one, we choose the median sentence query. The overview of the picked 4 queries are shown in Table 2.

### 3.4 Participants

Due to the high requirements of domain expertise in judicial judgment process, we only recruited participants with strong background in law. Specifically, we sent out invitations to law schools or firms and required participants to have a master’s degree in criminal law currently studying, master’s degree students expected to enter in the autumn of 2023, or doctoral degree students in criminal law. In total, we successfully recruited 94 law students as participants, more than 88% of whom have passed National Judicial Examination in China.

Behavior logs of the participants collect such as clicking buttons or taking breaks. We conducted a pilot study and found it impossible for a participant to finish the whole experiment (including four tasks and a questionnaire) within 30 minutes if they take it seriously. As shown in Figure 2, the averaged time to finish the experiment is about 75 minutes. Therefore, we filtered out data from 5 participants with finish time less than 30 minutes.

For each task (query case), we provide three types of retrieval results: the light group, control group, the harsh group. Specifically, we first assume that the top 30 cases for each query case provided in the LeCaRDv2 dataset are relevant cases, which is reasonable because LeCaRDv2 has retrieved and ranked cases using a combination of multiple SOTA legal retrieval models. Then, we sort these cases based on the severity of the penalty in their sentences in increasing order. We further split the list into three set according to the order: the top 10 is the light set (which contains low-penalty...
Table 1: Questionnaire content

<table>
<thead>
<tr>
<th>Question Topic</th>
<th>Question Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satisfaction</td>
<td>You feel satisfied with the overall experiment.</td>
</tr>
<tr>
<td>Confidence</td>
<td>You have confidence in your sentence.</td>
</tr>
<tr>
<td>Effort</td>
<td>The completion of the task requires a significant amount of mental effort (e.g., calculation, thinking, decision-making, and memory) [9].</td>
</tr>
<tr>
<td>Overall Unbiasedness</td>
<td>The cases retrieved provided for the task are fair.</td>
</tr>
<tr>
<td>Gender Unbiasedness</td>
<td>When providing the cases retrieved, the system tends to assign harsher sentences to male defendants (or female defendants).</td>
</tr>
<tr>
<td>Age Unbiasedness</td>
<td>When providing the cases retrieved, the system tends to assign harsher sentences to younger defendants (or older defendants).</td>
</tr>
<tr>
<td>Region Unbiasedness</td>
<td>When providing the cases retrieved, the system tends to assign harsher sentences to defendants from certain regions.</td>
</tr>
<tr>
<td>Usefulness</td>
<td>The cases retrieved assist you in completing the task.</td>
</tr>
</tbody>
</table>

* Answers are chosen from completely disagree, disagree, neutral, agree and completely agree.

Table 2: Overview of task cases

<table>
<thead>
<tr>
<th>Crime</th>
<th>Original Sentence (months)</th>
<th>Document Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Injury</td>
<td>7</td>
<td>1063 words</td>
</tr>
<tr>
<td>Robbery</td>
<td>42</td>
<td>1160 words</td>
</tr>
<tr>
<td>Rape</td>
<td>39</td>
<td>1333 words</td>
</tr>
<tr>
<td>Murder</td>
<td>156</td>
<td>4951 words</td>
</tr>
</tbody>
</table>

Table 3: Overview of different groups

<table>
<thead>
<tr>
<th>Group</th>
<th>Light set</th>
<th>Median set</th>
<th>Harsh set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Control</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Harsh</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Figure 2: Experiment finish time distribution

4 RESULTS

In order to address the research questions, we conduct a user study that aimed to examine the distribution of punishment across different groups. If significant differences are observed in the distribution of punishments across these groups, we would posit that the cases retrieved have an influence on users’ sentence. To analyze the data, we first employ Levene’s Test to assess the homogeneity of variance and subsequently use a two-sample t-test to determine significance across different groups.

To further bolster our findings, we utilize the results of the questionnaire to emphasize the validity of our experiment. If a majority of participants expresses high confidence in their sentences, it suggests that they are not aware of any manipulation and thus strengthens the credibility of our study. Similarly, if a majority reports that the cases retrieved are highly beneficial, this implies that their decision-making processes are influenced by the cases retrieved rather than solely relying on their knowledge and cognitive abilities, which further enhances the reliability of our experiment.

4.1 User Study Results

The sentence distribution of each task across different groups is shown in Figure 3 and the sentences’ statistics and significance are listed in Table 4.

Due to the fact that whether to reduce or not to reduce a punishment is within the discretion of judges in legal practice, in this experiment scenario, participants have made deviations within the scope of discretion due to the influence of cases retrieved. Moreover, the trend of deviation is consistent with that of participant grouping, indicating that the case retrieval system has influenced
the decision-making process of participants in judicial judgement. Specifically, if participants are given reference to harsher punishments, they tend to make harsher judicial judgements, which is significant. However, if participants are given reference to lighter punishments, their tendency is not significant.

In the task of intentional injury, it is observed that some participants in the experiment assigned the punishment of supervision or detention to the case details in the query. The number of cases with the punishments of supervision or detention is shown in Table 5. It can be found that in the light punishment group, no participant made the choice of judicial judgement with the punishment of supervision or detention. However, in the control group and harsh punishment group, there are participants who assigned the punishment of supervision or detention, not just an isolated case. The number of participants with such assignments is significant. Furthermore, participants in the harsh punishment group tend to assign the punishment of supervision or detention to judicial judgements. This is a phenomenon worth further study.

After completing the experiment, we conducted one-on-one interviews with all participants who had chosen to assign the punishment of supervision or detention. In the interviews, all ten participants stated that when conducting the experiment, they made judicial judgements with the punishment of supervision or detention influenced by cases retrieved. There are two thoughts among participant groups regarding this influence:

- The first is that participants believe that there are some similarities between the cases retrieved and the task case’s circumstances, according to which the sentence with similar circumstances should be assigned to the punishment of supervision or detention based on other circumstances that could reduce the punishment in the task case.
- The second is that participants believe that compared to the cases retrieved, the task case’s circumstances are significantly less severe and have a smaller social influence. However, similar circumstances in the cases retrieved are sentenced to less than one year in prison, so the punishment of supervision or detention should be assigned to the task case.

Regardless of which approach, it shows that participants received influence from cases retrieved when making judicial judgements. From the perspective of criminal law sentences, supervision and detention are the lightest sentences for intentional injury crimes. If participants claimed that they did not receive influence from cases retrieved and only relied on the circumstances of the task case to make such a sentence, then it may be true that they did not consider cases retrieved deeply enough. However, if participants claimed that they considered cases retrieved before making non-imprisonable judicial judgements, then it can be argued that they made a leap from “could be reduced” to “should be reduced” due to the mitigating circumstances being statutory “could be reduced” rather than “should be reduced” and the original sentence being seven months in jail as a suspended sentence. Therefore, in this scenario, it can be believed that a reference case retrieval system influenced participants’ decision-making process in judicial judgements.

4.2 Questionnaire Results
The results of the 5-level user questionnaire are shown in Figure 4. Most participants are generally satisfied with the experiment, had high confidence in their judicial judgement abilities, and found case studies to be helpful during the experimental process. Furthermore, since this experiment did not include any gender, age, or geographic bias in the case studies, most participants disagreed with the notion that such biases existed in the case studies, which is consistent with expectations.

Since most participants claim that the cases retrieved have assist their sentencing process, which supports the result of our user study, that is, shows that the sentences users make are influenced by our shown cases rather than decision-making by their knowledge and cognition only regardless of cases retrieved.

4.3 Data Access
The data used throughout the research and the user study results are available at: https://github.com/Benson0704/LegalJudgeUserStudy

5 CONCLUSIONS
Through statistical analysis of the experiment data, this study demonstrated that there is an influence of the case retrieval system on users. Specifically, the experiment showed that if participants are given reference to harsher sentences, they tended to make harsher judicial judgements, which is significant. However, if participants are given reference to lighter sentences, their tendency is not significant. Furthermore, through analyzing case-specific data from the experiment, it is found that the distribution of cases may affect users’ decision-making process. The direction and pattern of influence need to be analyzed specifically and may contradict intuition.

This study still has limitations, such as that sentencing itself is a highly technical, practical, and subjective activity. Most law students lack practical experience and have not undergone standardized training in sentencing. In addition, sentencing is also affected by a series of uncontrollable factors such as the economic situation of the jurisdiction where the judiciary is located, criminal policies, and overall sentencing balance within a school. These factors are far beyond the capabilities of current experiment participants, which can adversely affect the validity and objectivity of the experiment.

ACKNOWLEDGMENTS
This work is supported by Quan Cheng Laboratory (Grant No. QCLZD202301) and the Natural Science Foundation of China (Grant No. 62002194).

REFERENCES
Table 4: Sentence results of different groups

<table>
<thead>
<tr>
<th>Crime</th>
<th>Group</th>
<th>Participants</th>
<th>Median Sentence (months)</th>
<th>Average Sentence (months)</th>
<th>Significance ($P$ value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Injury</td>
<td>Light</td>
<td>27</td>
<td>7</td>
<td>6.74</td>
<td>0.0646 vs. Control</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>30</td>
<td>6</td>
<td>5.73</td>
<td>0.1293 vs. Harsh</td>
</tr>
<tr>
<td></td>
<td>Harsh</td>
<td>32</td>
<td>7</td>
<td>7.28</td>
<td>0.5567 vs. Light</td>
</tr>
<tr>
<td>Robbery</td>
<td>Light</td>
<td>30</td>
<td>24</td>
<td>32.93</td>
<td>0.4832 vs. Control</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>29</td>
<td>24</td>
<td>37.10</td>
<td>0.0241 vs. Harsh</td>
</tr>
<tr>
<td></td>
<td>Harsh</td>
<td>30</td>
<td>48</td>
<td>49.60</td>
<td>0.0000 vs. Light</td>
</tr>
<tr>
<td>Rape</td>
<td>Light</td>
<td>30</td>
<td>31</td>
<td>30.53</td>
<td>0.2052 vs. Control</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>31</td>
<td>34</td>
<td>34.00</td>
<td>0.0422 vs. Harsh</td>
</tr>
<tr>
<td></td>
<td>Harsh</td>
<td>28</td>
<td>39</td>
<td>38.68</td>
<td>0.0048 vs. Light</td>
</tr>
<tr>
<td>Murder</td>
<td>Light</td>
<td>30</td>
<td>144</td>
<td>136.53</td>
<td>0.1997 vs. Control</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>26</td>
<td>126</td>
<td>135.08</td>
<td>0.0081 vs. Harsh</td>
</tr>
<tr>
<td></td>
<td>Harsh</td>
<td>33</td>
<td>168</td>
<td>161.45</td>
<td>0.0440 vs. Light</td>
</tr>
</tbody>
</table>

Table 5: Supervision or detention sentences in injury

<table>
<thead>
<tr>
<th>Group</th>
<th>Supervision or Detention Sentences</th>
<th>Total Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>Control</td>
<td>4</td>
<td>30</td>
</tr>
<tr>
<td>Harsh</td>
<td>6</td>
<td>32</td>
</tr>
</tbody>
</table>

Figure 3: Punishment distribution of different groups

Figure 4: Questionnaire results


