Conversational vs Traditional: Comparing Search Behavior and Outcome in Legal Case Retrieval

Bulou Liu[1], Yueyue Wu[1], Yiqun Liu[1*], Fan Zhang[1], Yunqiu Shao[1]
Chenliang Li[2], Min Zhang[1], Shaoping Ma[1]
[1]Department of Computer Science and Technology, Institute for Artificial Intelligence,
Beijing National Research Center for Information Science and Technology, Tsinghua University
[2]School of Cyber Science and Engineering, Wuhan University
yiqunliu@tsinghua.edu.cn

ABSTRACT

In recent years, legal case retrieval has attracted much attention in the IR research community. It aims to retrieve supporting cases for a given query case and contributes to better legal systems. While using a legal case retrieval system, it's often difficult for users to construct accurate queries to express their information need, especially when they lack sufficient domain knowledge. Since conversational search has been widely recognized to fulfill users' complex and exploratory information need, we investigate whether conversational search paradigm can be adopted to improve users' legal case retrieval experience. We design a laboratory-based study to collect users' interaction behavior and explicit feedback signals while using traditional and agent-mediated conversational legal case retrieval systems. Based on the collected data, we compare search behavior and outcome of these two different kinds of interaction paradigms. Compared with the traditional one, experimental results show that users can achieve better retrieval performance with the conversational case retrieval system. Moreover, conversational system can also save users' efforts in formulating queries and examining results.

CCS CONCEPTS

• **Information systems** → *Collaborative search.*

KEYWORDS

Legal case retrieval, conversational search, user study

ACM Reference Format:
Bulou Liu[1], Yueyue Wu[1], Yiqun Liu[1*], Fan Zhang[1], Yunqiu Shao[1], Chenliang Li[2], Min Zhang[1], Shaoping Ma[1]. 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), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3463064

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

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463064

1 INTRODUCTION

In recent years, legal case retrieval has attracted much attention in the IR research community. It aims to retrieve supporting cases for a given query case and contributes to better legal systems. Along with statutes, prior cases decided in courts of law are primary legal materials in various law systems. For instance, prior cases are fundamental for a lawyer who is preparing the legal reasoning in the countries that follows the common law system. In countries following the civil law system, establishing legal search systems is also increasingly important because it promotes the consistency in application of law and the supervision on judges [9]. Existing works show that an automatic system not only performs the retrieval tasks with higher performance than lawyers, but it also finishes them more efficiently [13]. While using existing legal case retrieval systems, users usually issue queries to describe their information needs [7, 12]. However, users always feel difficult to construct sophisticated queries to express their information need accurately, especially when they lack sufficient domain knowledge [11, 19]. Therefore, the legal case retrieval process usually requires more effort to formulate queries and examine results than web search due to the lack of legal domain knowledge of ordinary users.

Conversational search is the embodiment of an iterative and interactive information retrieval system that proactively refines user requests and search results [15]. It can help user better express their information needs [15] and improve search accuracy during search sessions [5]. Therefore, conversational search has been widely recognized as a better choice to fulfill users' complex and exploratory information needs [4, 16]. Conversational search paradigm has been adopted in multiple search scenarios, such as web search [20], product search [21], academic search [3] and so on. However, it is still an open question whether conversational search paradigm helps legal case retrieval, which is a typical complex information search scenario.

In this paper, we investigate whether conversational search paradigm can be adopted to improve users' legal case retrieval experience. Specifically, we compare conversational legal case retrieval with traditional legal case retrieval to address the following research questions:

- **RQ1:** What are the differences in search interaction behavior between traditional and conversational legal case retrieval?

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

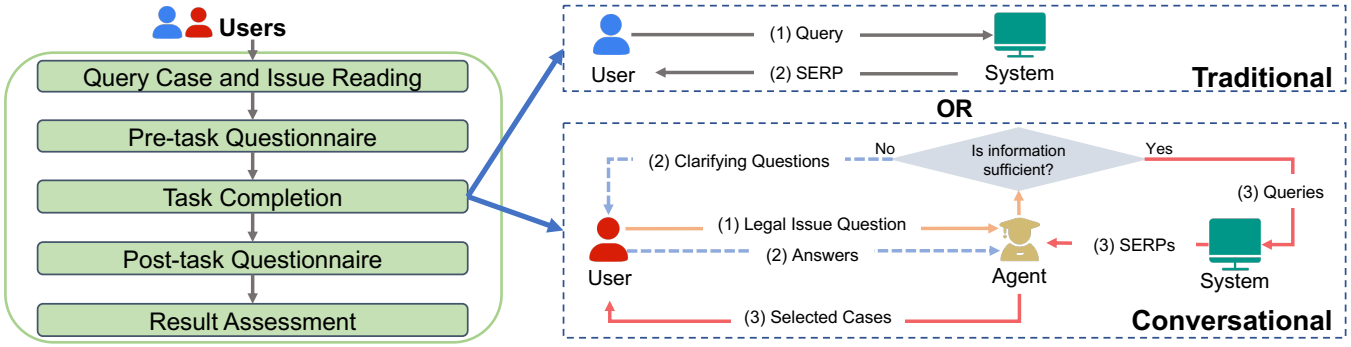

Figure 1: Data collection procedure.

- **RQ2:** What are the differences in user's search outcome between traditional and conversational legal case retrieval?

To shed light on these research questions, we first summarize the procedures of the two legal case retrieval paradigms. Then we conduct a lab-based user study to collect user behavior and explicit feedback signals using both traditional and agent-mediated conversational legal case retrieval systems. It's worth noting that no available conversational legal case retrieval system exists currently. Therefore, we recruit legal experts as intermediary agents to complete the procedure. Meanwhile, we also release the first conversational legal case retrieval dataset[1] with corresponding search behavior data to support the future research and promote the research of conversational search in other scenarios.

2 RELATED WORK

Legal case retrieval is a specialized IR task, which is different from ad-hoc retrieval in various aspects. Opijnen et al. [18] summarized the uniqueness of relevance in law from six dimensions. Turtle et al. [17] concluded the characteristics of legal documents, including professional legal expressions, the special logical structures and so on. However, the existing legal case retrieval systems still followed a traditional search paradigm in which users issued keyword-based queries to describe their information needs [7, 12]. Even the well-known legal case retrieval systems, e.g., Westlaw [6], require highly search skills and professional knowledge.

The conversational search paradigm has been investigated in various search scenarios. Zamani et al. [20] introduced Macaw, an open-source framework with a modular architecture for open domain conversational information seeking. Zhang et al. [21] proposed a multi-memory network to ask questions for improved e-commerce recommendation. Balog et al. [3] proposed to develop and operate a prototype system, Scholarly Conversational Assistant, that would serve as a useful academic search tool. Compared to these studies, our work is the first to investigate whether conversational search paradigm can be adopted to improve users' legal case retrieval experience.

3 DATA COLLECTION

To investigate the differences between traditional and conversational legal case retrieval, we conduct a lab-based user study with 55 search tasks. In this section, we describe the details of the user study (ref. Figure 1) and the dataset we collected.

[1]https://github.com/BulouLiu/Conversational-vs-Traditional-Legal-Case-Retrieval

Table 1: The statistics of the dataset in our user study.

	#Total	#Annotated	#Relevant	Fleiss's κ
Traditional	21.36	7.691	3.291	0.714
Conversational	13.85	4.782	2.364	0.658

3.1 Traditional vs Conversational

We first introduce the procedures of traditional and conversational legal case retrieval, which are shown in the right part of Figure 1.

3.1.1 Traditional Legal Case Retrieval. Though legal case searching is a more complex task, but most of the legal case retrieval systems are still keyword-based, i.e., taking the keywords issued by the users as inputs. We summarize the procedure as follows:

(1) The user submits a query to the legal case retrieval system.
(2) The system returns a search engine result page (SERP) to the user.

3.1.2 Conversational Legal Case Retrieval. As for conversational search paradigm, we add an intermediary agent to complete the procedure. The agent needs to understand users' intents via conversations, construct queries and pick cases from SERPs for the user. Specifically, the procedure contains the following steps:

(1) The user submits a legal issue question to the agent in natrual language.
(2) The agent asks clarifying questions [2] until the background information of the search issue is sufficient.
(3) The agent submits queries to the legal case retrieval system. She then selects cases from the SERPs and responds to the user with the selected ones.

In both the search paradigms, users examine the results and repeat the above steps until they find enough information or their patience is exhausted.

3.2 Tasks and Participants

We collected 55 search tasks from legal practitioners' real information need via online forums and social networks, including 20

Table 2: Comparison of search behavior. † indicates that the difference between the search paradigms is statistically significant at 0.05 level using Mann-Whitney U test.

Measure	Overall		In-domain		Out-domain	
	Traditional	Conversational	Traditional	Conversational	Traditional	Conversational
#query	3.836	1.364†	3.300	1.333†	4.143	1.387†
#case	7.727	2.145†	6.750	2.083†	8.286	2.194†
task time (s)	1371.8	546.1†	1124.6	573.2†	1513.1	525.1†
dwell time per case (s)	90.92	251.8†	91.52	276.5†	90.58	232.8†

civil (involving Civil Code), 21 criminal (involving Criminal Law) and 14 commercial (involving Company Law, Expertise Bankruptcy Law and Insurance Law) tasks. Each task contained a query case description and a legal issue. Users were expected to retrieve legal cases which may help to answer the issue question.

In our user study, there were two kinds of participants: users and agents. As for users, we recruited 110 participants (41 males and 69 females) via online forums and social networks, including lawyers, prosecutors and college law students. Each task was conducted by two different users in traditional and conversational legal case search system, respectively. This can avoid the effect of user's knowledge growth through the prior search session. We recruited 4 graduate students in law school (1 for civil law, 2 for criminal law and 1 for commercial law) to be agents. They only participated in the task related to their research fields, i.e., to ensure an adequate level of domain expertise. As for the legal case retrieval system, we choose a leading commercial legal search engine[2] in China. Users and agents had a conversation (just in text form) via Zoom[3].

3.3 Procedure

Before the experiments, we firstly requested each participant to complete a warm-up search task by their experimental search paradigm. The procedure of our user study is shown in the left part of Figure 1. We introduce the details as follows:

Query Case and Issue Reading. In the first step, the user read the query case description and the legal issue carefully. He or she could refer to the query case any time during searching.

Pre-task Questionnaire. Next, the user was asked to finish a pre-search questionnaire, including: domain knowledge level, task difficulty level, and interest level of the task with a 5-point Likert scale (1: not at all, 2: slightly, 3: somewhat, 4: moderately, 5: very).

Task Completion. After that, the user could perform searches by her experimental search paradigms (traditional or conversational). At this step, we collected the user's and agent's interactions (including queries, clicks and etc.) in traditional and conversational paradigm, respectively. Moreover, in conversational paradigm, we recorded the conversation contents, including search questions, clarifying questions, the cases returned by the agent.

Post-task Questionnaire. Once browsing the provided cases, the user was required to complete a post-task questionnaire. At this step, we collected explicit feedback signals on the experience of searching, including five-grade workload and satisfaction.

[2]https://ydzk.chineselaw.com/case
[3]https://zoom.us/

Result Assessment. After completing the post-task questionnaire, the user was asked to annotate the cases that he or she examined before. Each case was presented to the user again and he or she was asked to annotate relevance for each case. The relevant labels are on binary scale (1: irrelevant, 2: relevant).

3.4 Data Annotation

After collecting user search behavior and explicit feedback signals, we further recruited three additional legal experts (PhD in law, 2 for Civil and Commercial Law and 1 for Criminal Law) to annotate relevance for each case that users and agents clicked in traditional and conversational search paradigm, respectively, on binary scale. The Fleiss's κ among three assessors was 0.692, indicating substantial agreement [8]. If there were disagreements, we took the result of the majority vote. As for the cases that weren't clicked, we simply regarded them as irrelevant. Table 1 shows the quantity of cases that were returned by the legal case retrieval system (#Total), annotated by the assessors (#Annotated) and regarded as relevant (#Relevant) on average of each task.

4 DATA ANALYSIS

4.1 Comparison of Search Behavior

To address **RQ1**, we focus on four indicators that are highly related to users' search efforts. Specifically, we first compare the number of queries (**#query**) that users proactively submit and the number of cases (**#case**) that users examine between the two search paradigms. Note that in conversational legal case retrieval we regard the legal issue questions submitted by the users, rather than the queries issued by the agent, as the "queries" here. As shown in Table 2, users issue 3.836 queries and examine 7.727 cases on average in traditional legal case retrieval, which are significantly larger than those in the conversational search diagram (1.364 queries and 2.145 cases, $p < 0.05$ from the Mann-Whitney U test). These observations illustrate that conversational search paradigm can save users' efforts in formulating queries and examining results. Besides, we analyze users' **dwell time per case** and total **task time**. The results in Table 2 show that even though the total task time is longer in conversational legal case retrieval, users still spend more time on each examined case. It indicates that users examine each case more carefully and patiently in conversational search paradigm.

We also analyze search behavior difference between the two search paradigms with different level of domain knowledge. According to users' five-grade domain-knowledge level from pre-task questionnaire, we further divide users into in-domain group (4-5 points) and out-domain group (1-3 points). In traditional search

Table 3: Comparison of search outcome. The better results are highlighted in boldface. † indicates that the difference between the two search paradigms is statistically significant at 0.05 level using Mann-Whitney U test.

Group	Measure	Overall		In-domain		Out-domain	
		Traditional	Conversational	Traditional	Conversational	Traditional	Conversational
Subjective	Workload	3.055	**2.273**†	3.100	**2.083**†	3.029	**2.419**†
	Satisifaction	3.782	**4.255**†	4.100	**4.292**	3.600	**4.226**†
Objective	Success	0.691	**0.836**	**0.800**	0.792	0.629	**0.871**†
	Precision (all)	0.175	**0.197**	0.154	**0.182**	0.187	**0.209**
	Precision (last)	0.278	**0.390**†	0.262	**0.328**	0.287	**0.438**†
	Precision (visited)	0.525	**0.755**†	0.549	**0.774**	0.512	**0.740**†

paradigm, 20 users are in-domain and 35 users are out-domain. In conversational search paradigm, 24 users are in-domain and 31 users are out-domain. From Table 2, we find that the observations are consistent regardless whether users have sufficient domain knowledge or not.

Regarding **RQ1**, our finds are as follows: Regardless of domain knowledge level, 1) Conversational search paradigm can save users' efforts in formulating queries and examining results in legal case retrieval; 2) Users examine each case more patiently in conversational legal case retrieval than traditional legal case retrieval.

4.2 Comparison of Search Outcome

To address **RQ2**, We investigate search outcome via objective metrics and subjective feedback signals. Specifically, we compare the search output under the two search paradigms from multiple aspects, including perceived workload, user satisfaction, search success, query performance and accuracy of visited results. We also take the effects of domain knowledge level into consideration. The results are shown in Table 3. Our observations are as follows:

Perceived Workload. We evaluate users' workload through their perceived five-grade response. Users reported 3.055 and 2.273 scores of traditional and conversational search paradigm, respectively. The differences are also significant no matter users have sufficient domain knowledge or not. This indicates that users think they pay less effort in conversational legal case retrieval.

User Satisfaction. User satisfaction measures users' subjective feelings about their interactions with the system [10]. We collected the user subjective satisfaction in a 5-level scale. We can find that users achieve higher satisfaction in conversational legal case retrieval than traditional legal case retrieval, especially when users are out-domain.

Search Success. Search success measures the objective outcome of a search process [1, 14]. Specifically, we measure the proportion whether users have found at least one relevant case. In conversational search paradigm, users have found at least one relevant case in 83.6% tasks. In traditional search paradigm, users just found relevant cases in 69.1% tasks. This difference is more significant in the out-domain user group (62.9% vs 87.1%). It illustrates conversational search can improve search success in legal case retrieval when the user lacks sufficient knowledge.

Query Performance. It's vital to construct appropriate queries to better express information need in legal case retrieval. Specifically, we use two kinds of precision (the proportion of relevant cases), *Precision (all) and (last)*, to measure the query performance. They represent the precision of result lists according to all queries on average and the last queries, respectively. From Table 3, we can find *Precision (all)* and *Precision (last)* in conversational search paradigm are both higher than that in traditional search paradigm. The difference in *Precision (last)* is more significant, especially in the out-domain user group. Considering that legal case retrieval is an exploratory search task, users (or agents) may have to make several trials before submitting an appropriate query in the last round. It indicates that conversational paradigm can improve the query performance that better express users' information need in legal case retrieval.

Accuracy of Visited Results. We further analyze the accuracy of visited results by *Precision (visited)*, which represents the precision of the cases that users clicked, to measure search accuracy. We can find conversational search paradigm has a significantly higher *Precision (visited)* than traditional search paradigm . It illustrates that the conversational legal case retrieval system returns more precise and accurate results after more effective screening.

Summary. To answer **RQ2**, we summarize our findings as follows: 1) Users pay less workload subjectively and achieve higher satisfaction and success in conversational legal case retrieval than traditional legal case retrieval; 2) Specifically, users can express their information need more accurately and obtain more accurate search results in conversational legal case retrieval.

5 CONCLUSION

In this paper, we investigated whether conversational search paradigm can be adopted to improve users' legal case retrieval experience. Centered on the research questions, we conducted a user study to compare traditional and conversational legal case retrieval. We have obtained several interesting findings. As for search interaction behavior, users issue less queries and examine less cases in conversational legal case retrieval. As for user's search outcome, users achieve higher satisfaction and success in conversational legal case retrieval, especially they lack sufficient domain knowledge. Specifically, conversational search paradigm can generate more appropriate queries and return more precise results to improve search accuracy. Our results reveal the necessity of adopting conversational search paradigm to legal case retrieval. We also release the first conversational legal case retrieval dataset. As for future work, we plan to design automatic models for conversational legal case retrieval systems.

REFERENCES

[1] Mikhail Ageev, Qi Guo, Dmitry Lagun, and Eugene Agichtein. 2011. Find it if you can: a game for modeling different types of web search success using interaction data. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 345–354.

[2] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*. 475–484.

[3] Krisztian Balog, Lucie Flekova, Matthias Hagen, Rosie Jones, Martin Potthast, Filip Radlinski, Mark Sanderson, Svitlana Vakulenko, and Hamed Zamani. 2020. Common Conversational Community Prototype: Scholarly Conversational Assistant. *arXiv preprint arXiv:2001.06910* (2020).

[4] Nicholas J Belkin, Colleen Cool, Adelheit Stein, and Ulrich Thiel. 1995. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert systems with applications* 9, 3 (1995), 379–395.

[5] Ben Carterette, Evangelos Kanoulas, Mark M. Hall, and Paul D. Clough. 2014. Overview of the TREC 2014 Session Track. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014 (NIST Special Publication, Vol. 500-308)*.

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

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

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

[9] 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.

[10] Diane Kelly. 2009. *Methods for evaluating interactive information retrieval systems with users*. Now Publishers Inc.

[11] 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.

[12] John O McGinnis and Russell G Pearce. 2019. The great disruption: How machine intelligence will transform the role of lawyers in the delivery of legal services. *Actual Probs. Econ. & L.* (2019), 1230.

[13] John O McGinnis and Steven Wasick. 2014. Law's algorithm. *Fla. L. Rev.* 66 (2014), 991.

[14] Daan Odijk, Ryen W White, Ahmed Hassan Awadallah, and Susan T Dumais. 2015. Struggling and success in web search. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 1551–1560.

[15] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. 117–126.

[16] Paul Solomon. 1997. Conversation in information-seeking contexts: A test of an analytical framework. *Library & Information Science Research* 19, 3 (1997), 217–248.

[17] Howard Turtle. 1995. Text retrieval in the legal world. *Artificial Intelligence and Law* 3, 1 (1995), 5–54.

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

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

[20] Hamed Zamani and Nick Craswell. 2020. Macaw: An extensible conversational information seeking platform. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2193–2196.

[21] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*. 177–186.