Towards Legal Case Retrieval

Yunqiu Shao
DCST, BNRist, Tsinghua University
Beijing 100084, China
shaoyq18@mails.tsinghua.edu.cn

ABSTRACT

Precedents (prior cases decided in courts of law) are primary legal materials in both common and civil law systems. With the rapid growth of digitalized legal documents, it takes great efforts of legal practitioners to search for relevant cases. Previous study [1] found that legal researchers retrieved less than 20% of relevant documents when they believed they had found over 75% using Boolean techniques. Given this situation, an efficient system for legal case retrieval can be of great benefit and thus has drawn increasing attention in academic and industrial IR research.

Generally, legal case retrieval involves retrieving prior cases that should be "noticed" regarding a given query case, where "noticed" is a legal technical term denoting that a precedent is relevant and can support the decision of a query case. Legal case retrieval can be viewed as a specialized IR task but it differs from the traditional ad-hoc text retrieval in distinct aspects. Firstly, both the query and candidate cases involve extremely long and complex texts. Secondly, the concept of "relevance" in legal IR is beyond the general "topical relevance " and involves various dimensions [5]. Thirdly, collecting accurate relevance judgments is quite expensive since it requires expert knowledge, which makes it challenging to construct a large dataset, especially with accurate labels.

Our research focuses on the legal case retrieval scenario. The development of retrieval models always sits at the core of IR research. The first research question is proposed as:

- ***RQ1:*** *How can we automatically retrieve relevant prior cases given a query case?*

Following the framework of recent benchmarks [3], **RQ1** is investigated solely using the case texts. In the practice of legal case retrieval, however, the existing IR system usually assists users more interactively. Given this situation, we propose our second research question,

- ***RQ2:*** *How do legal practitioners search for supporting precedents?*

Relevance is a key notion in IR [4] while it is expensive to collect accurate relevance judgments in legal case retrieval since the legal relevance is complex and requires domain knowledge. We are inspired to take a deeper look at what makes a relevant legal case and propose the third research question,

- ***RQ3:*** *What are the criteria of relevance in legal case retrieval practice?*

Although a variety of retrieval models have been developed for ad-hoc text retrieval, including traditional bag-of-words models and neural models [6], they face great challenges resulting from the differences mentioned above. To address **RQ1**, we propose to model paragraph-level interactions via the pre-trained language model, BERT [2], and then obtain the document-level relevance by aggregating the interaction representations. Regarding **RQ2**, we plan to combine log analysis, user study, and interview strategies and expect to obtain a thorough understanding of user's search strategies and examination patterns in legal case retrieval. With regard to **RQ3**, we would like to collect explainable relevance judgments and investigate the relevance decision process by conducting an annotation user study.

With the three research questions answered, we would like to combine those findings and aim to improve the performance of the legal case retrieval system in the legal practice.

CCS CONCEPTS

• **Information systems → Information retrieval**; *Retrieval models and ranking*; *Users and interactive retrieval.*

KEYWORDS

Legal search; relevance; user behavior

ACM Reference Format:
Yunqiu Shao. 2020. Towards Legal Case Retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3397271.3401457

ACKNOWLEDGMENTS

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

REFERENCES

[1] David C Blair and Melvin E Maron. 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun. ACM* 28, 3 (1985), 289–299.
[2] 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).
[3] Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2019. COLIEE 2019 Overview. In *COLIEE'19*. 1–9.
[4] Tefko Saracevic. 1975. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for information science* 26, 6 (1975), 321–343.
[5] 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.
[6] Jun Xu, Xiangnan He, and Hang Li. 2018. Deep learning for matching in search and recommendation. In *SIGIR'18*. ACM, 1365–1368.

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07.
https://doi.org/10.1145/3397271.3401457