Quick Check: A Legal Research Recommendation System

Merine Thomas
Center for AI and Cognitive Computing
Thomson Reuters
Eagan, MN, USA
merine.thomas@tr.com

Thomas Vacek
Center for AI and Cognitive Computing
Thomson Reuters
Eagan, MN, USA
thomas.vacek@tr.com

Xin Shuai[*]
Wissee Inc.
Sammamish, WA, USA
shuaixin.david@gmail.com

Wenhui Liao[*]
Minneapolis, MN, USA
wendy.liao2009@gmail.com

George Sanchez
Center for AI and Cognitive Computing
Thomson Reuters
Eagan, MN, USA
george.sanchez@tr.com

Paras Sethia
Center for AI and Cognitive Computing
Thomson Reuters
Toronto, Canada
paras.sethia@tr.com

Don Teo
Center for AI and Cognitive Computing
Thomson Reuters
Toronto, Canada
don.teo@tr.com

Kanika Madan
Center for AI and Cognitive Computing
Thomson Reuters
Toronto, Canada
kanika.madan@tr.com

Tonya Custis[*]
Autodesk AI Lab
San Francisco, CA, USA
tonya.custis@autodesk.com

ABSTRACT

Finding relevant sources of law that discuss a specific legal issue and support a favorable decision is an onerous and time-consuming task for litigation attorneys. In this paper, we present Quick Check, a system that extracts the legal arguments from a user's brief and recommends highly relevant case law opinions. Using a combination of full-text search, citation network analysis, clickstream analysis, and a hierarchy of ranking models trained on a set of over 10K annotations, the system is able to effectively recommend cases that are similar in both legal issue and facts. Importantly, the system leverages a detailed legal taxonomy and an extensive body of editorial summaries of case law. We demonstrate how recommended cases from the system are surfaced through a user interface that enables a legal researcher to quickly determine the applicability of a case with respect to a given legal issue.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**; • **Computing methodologies → Information extraction**.

KEYWORDS

recommendation; learning to rank; legal research

ACM Reference Format:
Merine Thomas, Thomas Vacek, Xin Shuai, Wenhui Liao, George Sanchez, Paras Sethia, Don Teo, Kanika Madan, and Tonya Custis. 2020. Quick Check: A Legal Research Recommendation System. In *Proceedings of the 2020 Natural Legal Language Processing (NLLP) Workshop, 24 August 2020, San Diego, US.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

When preparing or reviewing a legal brief, litigation attorneys spend a significant amount of time searching for the most pertinent authority to bolster or refute a particular point of law. This involves sifting through a collection of millions of primary and secondary sources of law, as well as past briefs and memoranda. The task is particularly challenging given the need for high recall; an incomplete legal research process can potentially miss a highly relevant source of law that would adversely impact the litigation strategy.

Early work in document recommendation for legal research focused on the retrieval of relevant authority and briefs through a combination of explicit user query input and implicit user browsing behavior [1] or by attempting to cluster legal issues into broader topics [8]. In this paper, we present an approach that considers the task from a citation recommendation perspective [3, 5]. Our system, Quick Check, complements the legal research process by extracting the core legal arguments of interest directly from a user's input brief document and recommending relevant primary and secondary sources of law. In particular, the system leverages a combination of full-text search, citation network analysis, and clickstream analysis to surface highly relevant case law opinions. Importantly, apart from the user's brief, no other user interaction is required by the system to interpret the legal issues and locate relevant authority.

While the structure and formatting styles of legal briefs in the U. S. federal and state court systems will vary depending on the court level and jurisdiction, a typical document will include at least the following main sections (or the equivalents thereof): (1) an Introduction articulating the party's claim and relief sought, (2) a

[*]Work done while at Thomson Reuters.

Statement of Facts that summarize key factual elements at issue and the procedural history of the case, (3) an Argument section containing the legal issues at hand and related supporting facts, and (4) a Conclusion summarizing the main points and the specific relief sought. The Argument section is typically further divided into subsections, each discussing a particular legal issue. We refer to each subsection as an issue segment. The recommendation system we describe follows an issue-segment-centric approach; potentially relevant cases are mined and ranked with respect to a particular issue segment in the brief.

2 TRAINING DATA COLLECTION

The case ranking component of the system (Section 3.3) was trained on a large corpus of graded issue-segment-to-case pairs. The initial pairs were collected from a combination of manual curation by attorneys and an early prototype of the system, while the bulk of the dataset was collected from the output of successive improvements to the system. The quality of a recommended case was graded on a five-point Likert scale, reflecting the degree to which a case is relevant to the legal issue at hand. A recommendation with a rating of 4 or 5 is considered highly relevant, while one with a rating of 1 is considered irrelevant. In total, we collected over 10K graded pairs from attorney-editors for model training. The briefs were chosen to cover a variety of jurisdictions, practice areas, and motion types.

3 SYSTEM OVERVIEW

Figure 1 gives an overview of the Quick Check system architecture. The recommendation system consists of three primary stages: Document Structure Extraction, Candidate Case Discovery, and Case Ranking.

3.1 Document Structure Extraction

The first stage of the pipeline converts a user's uploaded brief document into HTML, which is used for all downstream document section parsing logic. Stylistic information contained in the HTML tags provide an obvious indication of section headings. Therefore, the system searches for the presence of a combination of bold, alignment, and heading elements. Of primary interest to the recommendation system is the accurate identification of the Argument section of a brief. Thus, a set of high-precision rules is applied against the extracted set of headings to capture the top-level Argument heading, which may include terms such as "Discussion", "Memorandum", or "Analysis". Subsection headings in the Argument section are identified through the presence of a numbering or word capitalization convention.

Each issue segment of the Argument section is a collection of paragraphs and citations describing a particular legal issue. We consider each issue segment in isolation when discovering and ranking candidate cases.

3.2 Candidate Case Discovery

Given an issue segment, the system first collects a large pool of potentially relevant cases. This is done using both search-based and citation-based document discovery mechanisms.

3.2.1 Search-engine-based Candidate Discovery. Each paragraph within a segment discusses a particular aspect of the legal issue at hand. For each of these paragraphs, we perform full-text search across a corpus of about 12M case law opinions using a proprietary search engine tuned for the legal domain. To increase the jurisdictional relevance of results, the search is restricted to a subset of jurisdictions based on the corresponding jurisdictions of the citations present within the segment or the rest of the brief.

In addition to the case law opinions themselves, we consider cases from a context-aware citation recommendation perspective [3, 6]. In particular, we leverage an index of pseudo-documents, each representing a case, constructed in the following manner. For a given case, we consider all cases and previously filed briefs in which a citation to the case is made. The sentence preceding the citation reference within the document is extracted and added to the pseudo-document corresponding to the case. Thus, a case's pseudo-document is an aggregate of all extracted reference texts and provides a representation of the legal context in which a case is cited. A set of full-text searches using the issue segment paragraphs is also performed over this index.

3.2.2 Citation-based Candidate Discovery. The set of case citations within an issue segment (hereafter referred to as input citations) gives a valuable characterization of the legal issue being discussed. The system leverages this citation "profile" to find potentially related cases through the following means:

- **Case and brief citation network**: The most directly related cases are those that are bibliographically coupled to the input citations (i.e. cases citing the same input citations). Similarly, a brief citation network is constructed by decomposing the corpus of past filed briefs into issue segments. We then consider all bibliographically coupled segments. For both the case and brief-issue-segment networks, we extract the set of other cases that are cited in the coupled case or issue segment as candidate recommendations.
- **Statutory annotations**: Statutory annotations provide concise summaries of important cases that have interpreted a statute or regulation. They are organized editorially in a hierarchy of procedural topics. Candidate recommendations are extracted by considering the cases that are found within the same procedural topic as an input citation.
- **Pinpoint headnotes**: An input citation will often be accompanied by a direct quote from the cited case or a page number pinpointing the relevant portion of the case. Moreover, a case will often have one or more editorial summaries, called headnotes, that highlight important points of law in the case. Headnotes contain reference links to the corresponding location within the case document where the point of law is discussed. Thus, one can correlate the input citation to one or more headnotes in the cited case based on a combination of the pinpoint information and headnote reference links[1]. This is useful because extensive editorial annotations exist that identify explicitly the point of law (i.e. headnote) for which a case is citing another case. Therefore, the system

[1] If more than one headnote is identified, the most relevant headnote is determined based on a combination of text similarity and topic similarity measures, the latter of which leverages a legal topic taxonomy.

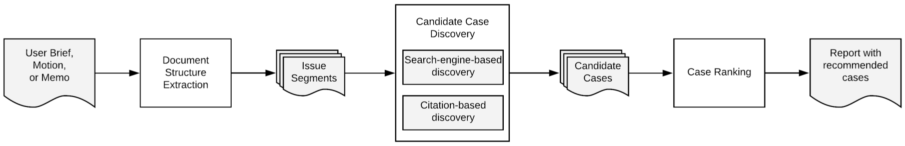

User Brief, Motion, or Memo
Document Structure Extraction
Issue Segments
Candidate Case Discovery
Search-engine-based discovery
Citation-based discovery
Candidate Cases
Case Ranking
Report with recommended cases

Figure 1: Overview of Quick Check system architecture.

is able to retrieve cases that cite the same case for the same reason as the input citation of the issue segment.

- **Clickstream analysis**: Within a particular research web session on our legal research platform, a user will interact with cases in a number of ways, including viewing the case, saving it to a folder, or printing the case document. Research session activity is aggregated across all users to provide implicit relevance feedback of cases. In particular, given the citation profile of the issue segment, the system finds cases that commonly appear within the same session.

3.3 Case Ranking

The pool of candidates collected from the discovery stage is passed through two ranking SVM models [7]. The first ranker uses metadata information corresponding to each of the discovery methods as features (e.g. how often the case was found in the top 5 results of searches, the number of input citations the case is bibliographically coupled with, etc.) and acts as a filter to reduce the pool size down to several hundred cases.

The second ranker leverages an additional set of features that measure the textual and topical similarity of the issue segment and the candidate case, where the issue segment is represented by either its textual content or the pinpoint headnotes of its input citations (Section 3.2.2). Textual similarity is computed using an edit-distance-based similarity measure, while topical similarity is assessed from the hierarchical similarity of the segment and candidate case when classified under a legal topic taxonomy using a legal topic classifier [1, 2]. Additionally, the recency of a case is taken into account at this stage.

Finally, the top-ranked candidates are fed to an ensemble-based pointwise ranker [4] leveraging additional features that analyze the results of the search-based discovery component. The model produces a probability score on the relevancy of a case, which is used to filter out poor quality recommendations prior to surfacing to the user.

4 RESULTS

The quality of the output recommendations is measured against a test set of nearly 500 briefs (corresponding to about 2K issue segments) using several metrics of varying granularity. Across all recommendations, the percentages of highly relevant, relevant, and irrelevant recommendations are 39%, 60.5%, and 0.5%, respectively. At an issue segment level, the percentages of segments with at least one highly relevant, at least one relevant or highly relevant, and at least one irrelevant recommendation are 67%, 97%, and 1%, respectively, while the mean *NDCG*@5 per issue of relevant or highly relevant recommendations is 0.66. For comparison, we note that the first ranker alone achieves a mean *NDCG*@5 of 0.62. Finally, at a brief level, the percentage of briefs where at least one-third of the recommendations are highly relevant is 55%.

5 DEMONSTRATION

Users can upload briefs that are in either an early draft or nearly completed state. They may also choose to analyze an old brief with potentially outdated authority or even an opposing party's document. When a brief document has been uploaded, the recommendation system pipeline is run. The entire pipeline completes in under a couple minutes for a brief document of typical length. The recommended cases are displayed and grouped by the corresponding issue segments. Each case is accompanied with additional information that helps to put the recommendation in context for the user, including the input citations that are related and the portion of text within the case found to be most similar to the issue segment. The latter is determined using a combination of legal topic classification (Section 3.3) and a vector space model representation of the issue segment and the recommended case. A recommendation may also be marked with additional tags highlighting if the case is from a high court, is frequently cited, or is less than 2 years old. Figure 2 shows the Quick Check interface for a sample brief.

After being presented with the recommended cases, a user may filter the the results based on the issue segment of interest, or by a specific date range or jurisdiction. The user can also choose to lower the threshold of the final ranker model to explore more recommendations from the system. Recommended cases can then be viewed in full or saved/downloaded for further review.

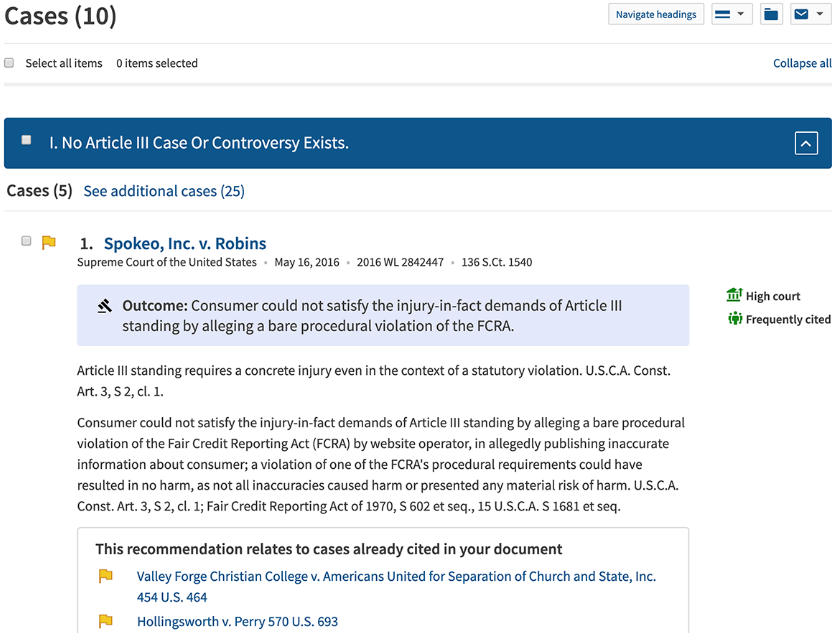
6 CONCLUSION

We presented Quick Check, a commercially available system that recommends cases with highly similar legal issues and facts given a user's input brief document. The system leverages a multitude of case discovery pathways and ranking models trained over a large annotated training set to extract the most relevant cases to a given legal issue. The system is robust against the wide variety of brief formatting styles and has been found to be effective across jurisdictions, practice areas, and motion types.

REFERENCES

[1] Khalid Al-Kofahi, Peter Jackson, M. Dahn, Charles Elberti, William Keenan, and John Duprey. 2007. A Document Recommendation System Blending Retrieval and Categorization Technologies. In *Proceedings of AAAI Workshop on Recommender Systems in e-Commerce*. 9–16.

[2] Khalid Al-Kofahi, Alex Tyrrell, Arun Vachher, Tim Travers, and Peter Jackson. 2001. Combining Multiple Classifiers for Text Categorization. In *Proceedings of the Tenth International Conference on Information and Knowledge Management* (Atlanta, Georgia, USA) *(CIKM '01)*. Association for Computing Machinery, New York, NY, USA, 97–104. https://doi.org/10.1145/502585.502603



Cases (10)

I. No Article III Case Or Controversy Exists.

Cases (5) See additional cases (25)

1. Spokeo, Inc. v. Robins

Supreme Court of the United States • May 16, 2016 • 2016 WL 2842447 • 136 S.Ct. 1540

Outcome: Consumer could not satisfy the injury-in-fact demands of Article III standing by alleging a bare procedural violation of the FCRA.

Article III standing requires a concrete injury even in the context of a statutory violation. U.S.C.A. Const. Art. 3, S 2, cl. 1.

Consumer could not satisfy the injury-in-fact demands of Article III standing by alleging a bare procedural violation of the Fair Credit Reporting Act (FCRA) by website operator, in allegedly publishing inaccurate information about consumer; a violation of one of the FCRA's procedural requirements could have resulted in no harm, as not all inaccuracies caused harm or presented any material risk of harm. U.S.C.A. Const. Art. 3, S 2, cl. 1; Fair Credit Reporting Act of 1970, S 602 et seq., 15 U.S.C.A. S 1681 et seq.

This recommendation relates to cases already cited in your document

Valley Forge Christian College v. Americans United for Separation of Church and State, Inc. 454 U.S. 464

Hollingsworth v. Perry 570 U.S. 693

Figure 2: Quick Check interface showing the top-ranked recommended case related to a legal issue extracted from the uploaded brief document.

[3] Michael Färber and Adam Jatowt. 2020. Citation Recommendation: Approaches and Datasets. *ArXiv* abs/2002.06961 (2020).

[4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2008. *The Elements of Statistical Learning* (second ed.). Springer New York Inc., New York, NY, USA, 339.

[5] Qi He, Daniel Kifer, Jian Pei, Prasenjit Mitra, and C. Lee Giles. 2011. Citation Recommendation without Author Supervision. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (Hong Kong, China) *(WSDM '11)*. Association for Computing Machinery, New York, NY, USA, 755–764. https://doi.org/10.1145/1935826.1935926

[6] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-Aware Citation Recommendation. In *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, North Carolina, USA) *(WWW '10)*. Association for Computing Machinery, New York, NY, USA, 421–430. https://doi.org/10.1145/1772690.1772734

[7] T. Joachims. 2002. Optimizing Search Engines Using Clickthrough Data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 133–142.

[8] Qiang Lu and Jack G. Conrad. 2012. Bringing Order to Legal Documents - An Issue-based Recommendation System Via Cluster Association. In *KEOD*.