Case law retrieval: accomplishments, problems, methods and evaluations in the past 30 years

DANIEL LOCKE, The University of Queensland

GUIDO ZUCCON, The University of Queensland

Case law retrieval is the retrieval of judicial decisions relevant to a legal question. Case law retrieval comprises a significant amount of a lawyer's time, and is important to ensure accurate advice and reduce workload. We survey methods for case law retrieval from the past 20 years and outline the problems and challenges facing evaluation of case law retrieval systems going forward. Limited published work has focused on improving ranking in ad-hoc case law retrieval. But there has been significant work in other areas of case law retrieval, and legal information retrieval generally. This is likely due to legal search providers being unwilling to give up the secrets of their success to competitors. Most evaluations of case law retrieval have been undertaken on small collections and focus on related tasks such as question-answer systems or recommender systems. Work has not focused on Cranfield style evaluations and baselines of methods for case law retrieval on publicly available test collections are not present. This presents a major challenge going forward. But there are reasons to question the extent of this problem, at least in a commercial setting. Without test collections to baseline approaches it cannot be known whether methods are promising. Works by commercial legal search providers show the effectiveness of natural language systems as well as query expansion for case law retrieval. Machine learning is being applied to more and more legal search tasks, and undoubtedly this represents the future of case law retrieval.

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

Additional Key Words and Phrases: Legal Information Retrieval, Case Law Retrieval

ACM Reference format:

Daniel Locke and Guido Zuccon. UNDER REVIEW. Case law retrieval: accomplishments, problems, methods and evaluations in the past 30 years. *ACM Comput. Surv.* 1, 1, Article 1 (UNDER REVIEW), 37 pages.

0360-0300/UNDER REVIEW/0-ART1 $15.00

https://doi.org/0000001.0000001

https://doi.org/0000001.0000001

1 INTRODUCTION

Law as a profession comprises a significant academic basis in that it focuses on finding, reading and analysing existing laws in various forms so as to advise a person on the legal bearing of their situation. Case law is one of the main sources of law in common law jurisdictions. This survey focuses on the task of finding relevant case law.

The focus of a lawyer's job on analysing case law comes about from the principle of *stare decisis* (doctrine of precedent). So that lawyers can advise clients, they must be able to find these binding legal principles.[1] The need and importance for lawyers to find these cases is so "[they] can discharge their duty owed to the court and to the administration of justice" [72]. Finding case law accounts for roughly 15 hours per week for a lawyer [67] or nearly 30% of their yearly working hours [89].[2]

Within the context of finding case law, the amount of information that must be searched is large, and it is growing rapidly. In the US, by 1962, 25,000 opinions were being published each year, and there were over 2.3 million decisions in the published case law reports [127]. By the early 2000s, this number had increased significantly; US appellate courts alone were handing down about 500 decisions each day [50]. Other legal sources also include large amounts of data: modern search providers search over a billion documents [2].

Throughout the last 20 years various methods have been identified as the standard for case law retrieval. In 2001, Moens [83] stated some form of concept based retrieval, relying on manual indexing of terms remained commonplace for case law retrieval. However, with the large number of cases being published, manual indexing was becoming ever more less feasible. She identified three models as standard for case law retrieval: Boolean retrieval, vector-space retrieval, and probabilistic retrieval. Maxwell and Schafer [79] identified similar models as the standard for case law retrieval in 2008, in addition to noting that knowledge engineering based methods for retrieval are common. Today, separate retrieval and ranking phases are the standard [24]. These systems score documents using multiple weighted features such as the court, user jurisdiction, citations and past

[1] *Stare decisis* "requires, broadly speaking, that like circumstances are considered in a like fashion; a case that considers a certain set of factual circumstances ... [is precedential] for any [analogous] future circumstances" and must be followed [72].

[2] This figure was for junior lawyers, with less than 10 years experience. Lawyers with more than 20 years experience spent on average 16% of their time researching. This may be due to experience or a change in role to a more managerial position. While more senior lawyers spend less time researching, their work is typically at a much higher cost to a client.

query logs [24, 69, 90]. Likewise, use of deep learning techniques in various stages of the retrieval process is becoming more common place, as we discuss throughout.

It has been over 20 years since Turtle [116] and nearly 20 since Moens [83] surveyed legal information retrieval. Yet in this time, little research has focused on ranking for case law retrieval as compared to information retrieval generally. Published research as to ranking for ad-hoc case law retrieval lags behind the general information retrieval community, despite legal information retrieval being one of the first applications of computers to information retrieval [12]. This is likely an incident of commercial search providers choosing not to publish any research. Aside from Moens [83] and Maxwell and Schafer's [79] identification of the standard methods for case law retrieval, there is only limited research on improving the effectiveness of these methods, or *a priori*, identifying or comparing their effectiveness. The majority of research in the legal information retrieval field has focused on argumentation retrieval, ontological frameworks and case-based reasoning, and technology-assisted review for discovery [5, 27].[3]

This survey focuses on the task of querying for ad-hoc[4] case law retrieval. But, legal information retrieval does not end there. Lawyers perform many search tasks. For instance, lawyers may search for legislation (laws enacted by parliament) (this may be ad-hoc search or by providing a factual scenario similar to ontological frameworks [124, 125]), civil codes (similar to legislation in a civil law jurisdiction) [48, 56–58], for documents in litigation, such as technology-assisted-review [5, 27], for patents, company information, dockets, or other documents generally in internal support system of a law firm [83]. Vast amounts of work have been done in these areas (for instance the TREC Legal tracks [5, 27] and ICAIL DESI series [4, 6]). Despite the existence of these other tasks, case law retrieval remains the sole focus of this article because of the unique challenges research in this area has encountered. Searching for case law is a peculiarity of common-law jurisdictions; the same need to find cases does not exist in a civil law jurisdiction.

By this article, a survey of work to date in the field of case law retrieval is examined. The article continues as follows: in Section 2 we outline the parts of a case; in Section 3 we describe in detail the nature of case law retrieval and provide a history of initial research in the area; in Section 4 we outline Cranfield style collections used in the evaluation of methods for case law retrieval; and, in Section 5 we outline various methods for case law retrieval as well as gaps in the literature and promising areas for future research.

[3] Discovery is the process by which parties to a dispute disclose all documents relevant to the issues between them (see rules 211 and 212 of the *Uniform Civil Procedure Rules 1999* (Qld).)

[4] Ad-hoc refers to the task of generic querying to satisfy an information need [28].

SUPREME COURT OF QUEENSLAND

CITATION:	*Drew v Makita (Australia) P/L* [2009] QCA 66
PARTIES:	**PAUL DREW** (plaintiff/respondent) **v** **MAKITA (AUSTRALIA) PTY LTD** ACN 001 117 335 (defendant/appellant)
FILE NO/S:	Appeal No 9558 of 2008 DC No 4664 of 2004
DIVISION:	Court of Appeal
PROCEEDING:	General Civil Appeal
ORIGINATING COURT:	District Court at Brisbane
DELIVERED ON:	24 March 2009
DELIVERED AT:	Brisbane
HEARING DATE:	12 March 2009
JUDGES:	Holmes and Muir JJA and Daubney J Separate reasons for judgment of each member of the court, each concurring as to the orders made
ORDERS:	**1. Appeal allowed.** **2. The orders made on 29 August 2008 are set aside.** **3. The matter is remitted to the District Court for a new trial on all issues other than quantum.** **4. The respondent's damages, should he be successful in establishing liability on the part of the appellant on the new trial, are as assessed by the primary judge in the reasons dated 29 August 2008.** **5. The costs of the first trial abide the result of the new trial.** **6. The respondent pay the appellant's costs of the appeal.**
CATCHWORDS:	APPEAL AND NEW TRIAL – APPEAL - GENERAL PRINCIPLES – RIGHT OF APPEAL – WHEN APPEAL LIES – ERROR OF LAW – PARTICULAR CASES INVOLVING ERRORS OF LAW – FAILURE TO GIVE REASONS FOR DECISION – ADEQUACY OF REASONS – where respondent claimed damages for injuries sustained by him when his left hand was severed by a circular saw – where respondent alleged that appellant manufactured or was deemed to have manufactured the circular saw and that a

Fig. 1. Portion of a headnote of a decision of the Queensland Court of Appeal: *Drew v Makita (Australia) Pty Ltd* [2009] QCA 66.

2 OVERVIEW OF A CASE

In Figure 1 we show a headnote of an example case, and in Figure 2 we show the start of the text of the same case's reasons. The headnote is not part of the decision and contains only metadata about the case. [5] Nonetheless, information in the headnote is relevant for filtering and searching on cases. The citation, and file numbers (FILE NO/S) may be searched on in the case of known-item retrieval. Likewise, the division, proceeding, delivery date, or judges may all be used for filtering results given an ad-hoc search. Typically the catchwords would be the only field outside of the text of the judgment that will be searched on. These catchwords are manually identified descriptions of the case provided by judges or court staff, and will typically include a list of cited cases and legislation.

[5] This differs to an American lawyers' understanding of a headnote, which refers to editorial annotations that summarise one point of law in a case.

[1] **HOLMES JA:** I agree with the reasons of Muir JA and with the orders his Honour proposes.

[2] **MUIR JA: Introduction**
On 29 August 2008, the plaintiff/respondent was awarded judgment against the defendant/appellant in the sum of $194,454 after a trial in the District Court for the respondent's claim for damages for injuries sustained by him on 19 December 2001 when his left hand was severed by a circular saw he was using. The respondent

Fig. 2. The beginning of the reasons for decision of *Drew v Makita (Australia) Pty Ltd* [2009] QCA 66

3 THE TASK OF CASE LAW RETRIEVAL, AND AN HISTORY

Within the task of case law search, the following tasks can be identified: (i) general ad-hoc querying; (ii) question answering; (iii) routing or selective dissemination; (iv) known item search; (v) navigation; and (vi) intra-document retrieval [115]. Querying has been said to pose the greatest problem for lawyers, in that they must: "(1) know and be able to articulate their information need; (2) know the content and storage structure of the documents in the database; and (3) be aware of the operation of the retrieval mechanism." [78]. While this survey focuses on the task of ad-hoc querying for case law retrieval, there will necessarily be overlap between the applicability of methods and concepts and a discussion of their results with the other tasks listed above.

In 2010, Bing [12] documented the history of case law retrieval from a practical perspective. This was less focused on any particular commercial search provider's system but with the early genesis of case law retrieval. In 1995, Turtle [116], on the other hand, provided more of an academic view of the history to date of case law retrieval. Given Turtle's work at Westlaw, his work is more detailed as regards their methods. Apart from the short summary below, we do not endeavour to repeat either of their syntheses. Rather, we focus on the developments made in the last 20 years. Turtle highlights the rapid developments of case law retrieval from the 1970s through to the 1990s. He states that prior to the 1970s "virtually all legal research was based on printed materials." At that point in time, major electronic providers only provided first-order logic based retrieval. But, by the early 1990s ordinary language systems had begun to emerge. Turtle [116] and Moens [83] discussed a combination of the following points of case law retrieval, which we further elaborate on:

(1) *the collections* that must be searched on are large (at the time of Turtle's [116] review, US legal text comprised about 50 gigabytes of data, and was growing at a rate of 2 gigabytes per year). A subset of all legal documents, being those that are publicly available in the US consists of approximately 240 gigabytes of data. This comprises roughly 4 million documents. By comparison, in 2012, commercial systems were searching on over 11 million cases [74];

(2) *the document length* is greater than those in other areas such as web search, but also varies widely. As an extreme example, the decision of *Bell Group Ltd (in liq) v Westpac Banking Corporation* (2008) 39 WAR 1 is 961 pages long, and at the other end of the spectrum, the decision of *Western Export Services Inc v Jireh International Pty Ltd* (2011) 282 ALR 604 is 6 paragraphs long. One must add that in the context of ad-hoc case law retrieval, the full text, or the catchwords, as a summary of the case, are the only fields that will frequently be searched on. The title is not descriptive of the issue discussed, and represents only the parties involved in a dispute; as such, the title is only relevant for known item retrieval. The full text rather than a summary or metadata is the most important aspect [121]. Further, not all documents will have a headnote summarising the issues. In this respect, headnotes are expert compiled either by an editor or by the respective decision maker and may be of varying quality. But, searching on headnotes represents one aspect in which search may progress. More modern controls over the publishing of judicial decisions has led to greater availability of headnotes.

(3) *there is significant added editorial value* with indexing, summaries, notes, and classification codes. Commercial services add summaries of decisions which may also be searched on. They may also classify decisions according to their topic or topics thereby aiding in retrieval (for example, Westlaw Key numbers[6]);

(4) *the relationship of the institution* creating the legal document is important (which court makes the decision has a bearing on its importance). Such an approach has been considered in early citation analysis approaches to find decisions relevant to a seed decision [112]. Tapper considered the differing jurisdiction of a decision in seeking to give a numeric value to a citation for the purpose of clustering. This sought to give an inverse value to a decision based on the level at which it was decided, with a higher court at the top, such as the Supreme Court of the United States, or the High Court of Australia receiving a lower value, and a lower court, such as an intermediate appellate court receiving a higher value. While Turtle does not make the point, so too may the specific author be important. Westlaw accords some weight to the court and user's jurisdiction in ranking case law [24]. For instance, Ravel Law[7] allows searchers to explore the specific decisions of a particular judge to analyse [1]. In fact, it premises a large number of its features on this ability.

[6] See legal.thomsonreuters.com/en/insights/articles/using-the-west-key-numbers-system

[7] A recently created commercial search engine: ravellaw.com

(5) *legal databases are the law* and while in other fields, old documents may not reflect current state of doing things (such as in evidence based medicine), cases from 200 or 400 years prior may still be applicable or informative;

(6) *the language* is unique. Judicial language has been said to be rich and feature very domain specific terminology [121], to feature no redundancy [103] and that it is not a pure sublanguage [83]. One thing is certain: legal sentences are frequently long and contain many clauses;

(7) *citations* are very important in legal decisions. The doctrine of precedent requires that courts that sit lower in the hierarchy must follow the decisions of higher courts. Also, judges must explain their reasons for a decision. To show that judges are conforming to binding authority they can explain their approach and explain how it accords within previous decisions. Hence, there is a rich network of citations. But, their effectiveness in aiding retrieval is not explored;

(8) *the audience* is wide, and ranges from laypersons to academics and lawyers [121];

(9) *the cost* of legal research is high. Subscriptions to commercial systems cost individuals thousands of dollars each year. Lawyers charge fees for their services, usually based on the time taken. Accordingly, minimising the time taken researching a legal question is key to reducing end costs for clients and making the legal system more accessible and affordable.

3.1 What is the nature of case law retrieval?

The nature of case law retrieval is an important question. It determines whether methods or systems for retrieval are effective. Maxwell and Schafer [79] summarised the early positions of Dabney [32, 33] that total recall is required. This position is understandable. It takes the view that as lawyers are responsible for their work they must be fully informed otherwise they face adverse cost awards or ethical breaches. This is the view taken by Lu and Conrad [73]. Such a position is common to all professional tasks (e.g. technology assisted review) [94, 109]. But legal work is frequently cost and time constrained. Gerson [41] suggested, however, that the task is merely one that is precision orientated. This is because of the amount of decisions a lawyer will be willing to read within a reasonable time and budget, and because of the existence of other tools such as those that allow a lawyer to follow the citation history of a case. He stated that because of the nature of legal information, where cases are processed and summarised by legal publishers, the key to legal research is to find one relevant decision quickly, from which other tools can be used to find other relevant decisions. Therefore, a legal information system should be judged by how quickly it can find relevant decisions [41]. This is certainly an attractive approach to legal research and it is one

Collection	Documents	Queries	Assessments	Purpose
Locke and Zuccon [71]	3,597,230	12	2572	general use
Locke et al. [72]	63,916	100	2645	query generation
Koniaris et al. [60]	63,742	330	105,036	diversification
Koniaris et al. [61]	3,890	330	unknown	diversification

Table 1. Collections previously created for use in case law retrieval.

where tools, such as Zhang and Koppaka [132], proposed and those enabling searchers to trace citations by paragraph,[8] so they can focus on the distinct legal question relevant, gain use.

Gerson [41] shared the view that the task is a recall orientated task. He places an important proviso on this view: that it is also one that requires maximisation of precision. The maximisation of precision is necessary to: (i) ensure that lawyers are not overburdened with information; and (ii) reduce costs associated with legal research and the provision of legal services. It is this view that the authors share. This view is sensible in light of the costs of legal services, whilst ensuring that quality is maintained. In this sense, case law retrieval is different to total recall tasks such as prior-art search in patent retrieval [104] or systematic review literature searching [49]. Lu and Conrad [73] appear to share a similar view of case law search as a high recall task that is cost constrained. Case law may, more appropriately, be considered interactive.[9] This has been long recognised [47]. Turtle [115] for instance, identified case law retrieval as an interactive task, with searchers willing to only look at somewhere around the top 20 results. The authors share this view. Recent discussion of queries used in commercial case law retrieval systems may support this [105].

4 AVAILABLE COLLECTIONS FOR CASE LAW RETRIEVAL

Test collection based evaluation of information retrieval has represented the standard since the 1950s [95, 123]. Despite this, until recently, there has been no standard collection of general utility for evaluating case law retrieval. And in this regard, work is still required. In Table 1, we outline current available collections for case law retrieval.

Locke and Zuccon [71] summarized the collections available for case law retrieval. Aside the 2017 collection of Koniaris et al. [61] which contains under 4,000 documents, Locke et al. [72] and

[8] See for example jade.io.

[9] E-discovery may also be considered highly interactive, with human in the loop processes [45].

Koniaris et al. [60] both created collections based on 60,000 United States Supreme Court cases. The collections of Locke et al. [72] and Koniaris et al. [60] contain the same documents.

Locke and Zuccon [71] have previously said that these collections (and the smaller collection of Koniaris et al. [61] to a greater extent) are not representative of how many cases a lawyer may realistically search on. This is because the current total number of legal decisions in common publicly accessible legal search systems range from about 850 thousand cases (Courtlistener) to over 4 million (Austlii). Comparatively, Westlaw[10] in 2012 reportedly contained 11 million documents that could be searched on [74] and Bloomberg Law presently contains around 1 billion documents [2].[11] Searching over such a small number of cases may be representative of searching a singular court or jurisdiction's decisions.

Further, as Locke and Zuccon [71] identified, the collections created by Locke et al. [72] (AIRS) and that of Koniaris et al. [60] are not appropriate for generally evaluating case law retrieval in so far as the assessments are concerned. The collection of Locke et al. [72] has an average of 26 assessments per topic. While, as Locke and Zuccon noted [71], a small number of relevance judgments is not appropriate for case law retrieval, as a recall orientated task [41], this was not a limitation of their work because they pooled methods prior to assessment. It does mean, however, that the collection cannot be reliably used to evaluate methods that were not pooled. The collection of Locke and Zuccon [71] addresses this issue. However, its small number of queries is problematic. Their collection, which comprises more than 3.5 million documents, contains a minimum of 200 assessments per topic, pooled through a number of standard queries, and expert user issued queries. The methods for the creation of this collection follow similar methods used to create internal collections at Westlaw [115].

Koniaris et al.'s [60] collection does not have a small amount of assessments. However, the assessments automatically gathered (they did not obtain expert relevance assessments). They took the highest ranked documents from an LDA topic model and labelled these documents as relevant. Looking at the topics used by Koniaris to generate the assessments also highlights the lack of general utility of the collection. The queries are a subset of the areas of law in Westlaw's Digest.[12] They represent a general search need rather than a specific information need. This is distinct from the problem that queries commonly issued in Lexis are very short and have been said to be

[10] www.westlaw.com, A popular commercial case law search system.

[11] It is not known what percentage of these are cases as opposed to other document types.

[12] A keynumber system of categorised areas and subareas of law. Areas of law can be searched or browsed by number.

insufficient to describe an information need [105]. In Table 2, we extract Locke and Zuccon's [71] summary of the queries used by Koniaris et al. [60] are compared with those of Turtle [115], which represent queries created by an expert searcher (a lawyer), rather than extracted from the Westlaw Digest, as Koniaris et al. [60]. This shows the artificial nature and the broad scope of the queries from Koniaris et al.. The queries employed by Koniaris et al. [60] may be more reflective of a general search to gain background information, rather than to search for a specific legal issue.

It is important to note that assessments for case law search collections are difficult to obtain. Relevance assessments based on citations are not appropriate. This is because a decision may cite another decision for any number of reasons, for any number of different topics. Tapper [111] views objective relevance assessments in the context of Cranfield based evaluation of retrieval systems as preferable: that being to take the citations in a case as relevance assessments rather than asking searchers for subjective assessments. But, he articulates two reasons why this is not appropriate however. First, that the citations may not be adequate. And second, that the citations may not be exhaustive. This second reason is likely in jurisdictions such as Australia and the UK, where citing every decision that considers the same question is cautioned against [34].

Finally, while Cranfield style collections were, and still are undoubtedly important in empirical analysis of the effectiveness of a retrieval system, the lack of large collections for case law retrieval may only pose a problem academically and not for commercial search providers. Large commercial legal search providers can gain vast amounts of user data, with huge amounts of queries being issued [105]. Vast query logs may mean that serving similar cases to users with similar queries may take precedence to optimizing effectiveness measures in a Cranfield style evaluation. This has definitely been an approach taken by Westlaw, where query logs are employed to generate surrogate documents from which features for a ranking functions can be learned [69].

Nonetheless, Cranfield style evaluations are still important, both where user data is not available, to ensure that a potential ranking method does not degrade search effectiveness too detrimentally, and for research, where in our view user-logs are not possible or limited largely. We view user-logs as infeasible in a research setting for several reasons. This relates to the sensitivity and confidentiality of legal searches. They are specific to advice to be given to, or cases being prepared for, a client. As such, without clients' consent, queries should not be contained in published query logs used for research. On the other hand, traditional Cranfield style collections can be more appropriately created. This is because, topics or queries can be expressed in an abstract fashion, without need for explicit approval from a client; a query or topic in and of itself does not give away anything

about a client or their case and as such there is no issue of confidentiality. The impact of limited user data may be minimal however, given the short time required to obtain large amounts of data, and also where methods to generate synthetic user data are used [90]. Creating such collections will undoubtedly be very costly. Cases are typically long and legal issues are complex. As such, in our view, lawyers will be necessary to assess relevance, as is done by commercial providers, who have in house teams of lawyers that perform these tasks.

While not specifically on the task of ad-hoc retrieval, several collections exist for closely related tasks. More recent COLIEE competitions include collections for finding cases relevant to a seed case, as well as identifying entailment between cited cases (how a cited case is treated) [51, 91].

Source	Query	Generation Method
Koniaris et al. [60]	Products Liability	Topic in Westlaw Laws of America Digest
Turtle [115]	(741 +3 824) FACTOR ELEMENT STATUS FACT /P VESSEL SHIP BOAT /p (46 +3 688) "JONES ACT" /P INJUR! /S SEAMAN CREWMAN WORKER	Manually created by expert searcher
Turtle [115]	What factors are important in determining what constitutes a vessel for purposes of determining liability of a vessel owner for injuries to a seaman under the Jones Act (46 USC 688)?	Natural language issue statement
Locke et al. [72]	"sovereign immunity" AND (immunity OR indemnif!) AND state AND suit AND (surrend! OR exist!) AND (tribe OR tribal OR "indian trib!")	Manually created by expert searcher

Table 2. Locke and Zuccon's [71] summary of queries previously employed in legal information retrieval studies. Koniaris et al.'s [60] queries were artificially created from Westlaw Digest topics. Turtle's [115] queries were created by lawyers. "/P" is a proximity based Boolean "and" operator where the operands must occur within the same paragraph. Natural language queries from Turtle's work are included for comparison.

5 METHODS FOR CASE LAW RETRIEVAL

In Table 3, we summarise the main methods for case law retrieval as well as the main works for each method. As briefly touched on above, several methods have, in the past, been identified as

Method	Section	Main works examined
Boolean and natural language	5.1	[32, 41, 115]
Conceptual search and case-based retrieval	5.2	[77, 118]
Question answering	5.3	[35, 88]
Query expansion	5.4	[15, 30, 100]
Query reduction	5.5	[70–72]
Search Result Diversification	5.6	[61]
Use of citation networks	5.7	[92, 126]
Deeper understanding of texts	5.8	[19, 43, 85, 106]

Table 3. Summary of methods and main works examined in this survey that comprise these methods.

the standard for case law retrieval. In the early 2000s, Moens [83] identified some form of concept based retrieval, relying on manual indexing of terms as commonplace, with three models as being standard for case law retrieval: Boolean retrieval, vector space retrieval, and probabilistic retrieval. She reported that legal text retrieval is primarily based on concepts identified by domain experts rather than on full text retrieval, in which indexing was frequently manually done and hence is very time consuming and expensive. While indexing of terms may be automated, she identified several reasons for why this leads to effectiveness of retrieval systems that are not comparable to systems where terms are manually indexed. Maxwell and Schafer [79] stated that there are two main approaches to legal information retrieval: (i) knowledge engineering; and (ii) natural language processing techniques.

There has been no comparison of these methods for case law retrieval. In fact, much of the research comparing, empirically, the effectiveness of case law retrieval systems as summarised below deals with evaluating commercial search systems. This is a significant gap in the literature. Attempts have been made to address this deficiency in published research in the AI and law community. Conrad and Zeleznikow [25, 26] have summarised the presence of evaluation in works in ICAIL. In ICAIL, the majority of presented works were theoretical. Despite this, works presenting algorithms are more frequently being evaluated empirically. While the lack of empirical evaluation of many methods discussed in this survey is problematic, these methods are still of interest. Not least for the intuitions and ideas behind their development. All that is further required is a more grounded analysis such as comparisons with baselines, human evaluation and statistical significance testing [26].

Review	System	Year	Query type	Recall	Precision	MR
Dabney [32]	Westlaw	1986	Boolean	0.197	0.269	
	LexisNexis		Boolean	0.114	0.261	
Dabney [33]	Westlaw	1993	Boolean	0.322	0.124	
	LexisNexis		Boolean	0.264	0.115	
Gerson [41]	Westlaw	1999	Natural language	0.31		2.3
	LexisNexis		Natural language	0.37		2.5
Mason [76]	Westlaw	2006	Unknown	-	0.810	
	LexisNexis		Unknown	-	0.740	

Table 4. Evaluations of the effectiveness of commercial search systems for ad-hoc case law retrieval. Note that each work used different collections of documents, queries and relevance assessments. The only evaluation that reports the particular version of search system used is Gerson, using WIN and FREESTYLE respectively.

A common starting point in any discussion of case law retrieval is the study by Blair and Maron [14]. But this is not a study considering case law retrieval – it considered retrieval of documents in a discovery setting. Similarities between other legal search tasks may nonetheless be present. The study analysed the effectiveness of a commercial litigation support system — an early precursor to technology assisted review for discovery. It involved the manual review of 40,000 documents. The study found that while users thought they had retrieved at least 75% of relevant documents, they had retrieved close to 20%. While this system did not evaluate retrieval of case law, it has been suggested that these results accurately reflect the effectiveness of case law retrieval systems [32]. But then again, extrapolating these results to case law retrieval systems has been cautioned against [44]. Blair and Maron hypothesised that the poor recall of the system in question may have been a result of semantic differences in the texts. Whether this is a problem is not known. Other external[13] studies of the effectiveness of commercial systems are summarised in Table 4.

Dabney's [32] analysis has little detail by way of the size of the collections and the queries used for evaluation. His subsequent work [33], however, detailed use of 23 areas of law summarised by practitioners (including a comprehensive set of relevant cases for that area of law). While the

[13] By external, we mean that they are performed on commercial systems by persons not employed by the organisation.

size of one collection is not known, the other contained over 1 million documents. Gerson [41] evaluated the effectiveness of the systems using 22 queries, measuring the greater of precision or recall at 20 documents. Mason [76] evaluated the effectiveness of the two systems using the same collection of documents. The collection's size is not detailed. Mason judged the first 10 results from 50 queries according to the measures adopted by TREC. The way in which Mason [76] determined relevance of a judgment to a query differs from the way adopted by Gerson [41] and Dabney [33], who both used annotated areas of law: a legal topic and list of relevant cases as determined by lawyers as an exhaustive list of relevant documents. Mason [76], on the other hand, adopted a subjective determination of the relevance of the first 10 retrieved documents, classifying returned documents as relevant or not-relevant.

In earlier work Dabney [32] suggested four potential ways in which the effectiveness of a system could be improved, including: (i) the ability of a system to search for plurals or lesser common synonyms (he also noted, without detail, that Lexis and Westlaw both search for plurals); (ii) the provision of more help to a user in "choosing search elements"; (iii) moving away from strict Boolean rules for determining whether a document is relevant or not; and (iv) usage of citations.

The substantial differences in the recall and precision in these previous studies may be due to increases in the effectiveness across these commercial search systems. Or it may be put down to the lack of control as regards the queries, collection, relevance assessments and methods used to evaluate these systems in these studies.

Another comparison of methods to note is that of the FIRE IrLed task [64]. This task involved retrieval of cited decisions from prior decisions where the citation name had been removed. While one will immediately note that this is not the exact task of ad-hoc case law retrieval, it does provide some examination of methods. There were 200 cases for which citations were removed, and 2000 prior cases from which these citations were to be identified. The problem may be viewed as an identification problem: finding the correct citation for each occurrence within a document. However, the task became a ranking problem by virtue of the measures used, these being *MAP*, *P*@10, *Recall*@100 and *MRR*. A range of methods were considered. These included query reduction using statistical based term selection [70], a language model using Dirichlet prior smoothing, probabilistic searching using BM25, a vector space model [113], ranking documents by the sum of scores from identifying the citation of legislative texts through regular expressions and returning prior decisions that also feature the same citations of articles, word distribution from an LDA topic model, and Doc2Vec [68] cosine similarity [64]. Similar to the IrLed tasks, is FIREs AILA track [10].

Work	Summary	Evaluation	Findings
[115]	Comparison of Boolean and natural language search	Cranfield style but non public collections	Natural language outperforms Boolean search
[32, 33]	Comparison of Boolean queries on two commercial systems	Non-typical evaluation on commercial system using small number of queries	LexisNexis outperforms Westlaw in Recall and Precision
[76]	Comparison of natural language queries on two commercial systems	Non-typical evaluation on commercial system using small number of queries	Westlaw outperforms LexisNexis in Precision
[41]	Comparison of natural language queries on two commercial systems	Non-typical evaluation on commercial system using small number of queries	LexisNexis outperforms Westlaw for combination of Recall and Precision
[79]	Summarises knowledge-base and natural language case law search	N/A	N/A

Table 5. Summary of works comparing Boolean and natural language search.

One of the AILA tasks is finding relevant cases given a factual scenario expressed in natural language. Again, this task is slightly different to that of ad-hoc case law retrieval. Bhattacharya et al. [10] summarise the methods attempted in the AILA track.

5.1 Boolean and natural language systems

Boolean retrieval represents the early foundation of case law retrieval systems. We summarise works that have investigated Boolean systems in Table 5.

In our view, the most comprehensive evaluation of the effectiveness of Boolean methods for case law retrieval is that by Turtle [115]. Turtle, then a researcher at Westlaw, compared the effectiveness of Boolean and natural language queries on two collections of 12,000 and 410,000 documents,[14] using Westlaw's case law retrieval system. This was an implementation of Turtle's inference network retrieval system [114, 117]. Evaluating the system's effectiveness over 44 queries, he concluded that evaluation on a commercial system shows natural language queries offered superior effectiveness to Boolean queries. On the smaller collection, his evaluation found a precision at 20 (P@20) and a recall of 0.423 and 0.244 for Boolean queries, and 0.57 and 0.329, respectively, for natural language queries. On the larger collection, his evaluation found a P@20 and a recall

[14] These collections were data internally available at Westlaw, and are used in their commercial system. We do not summarise these collections above, however the smaller is notable for its complete relevance assessements.

of 0.611 and 0.217 for Boolean queries and 0.759 and 0.269 for natural language queries. The evaluation, at what may be suggested to be a low depth for a recall orientated task - P@20, takes its motivation from Turtle's suggestion that lawyers are not willing to search through all results, but will evaluate only the first 20 or so before choosing to reformulate their query.

This remains the best, and in essence, the only study in the domain (aside from those discussed above that performed studies external to the system) as to both the comparative effectiveness of Boolean retrieval systems, and also of natural language systems. One criticism that may be made of the work is the method used to create Boolean queries compared to the natural language query. The natural language queries were fixed. On the other hand, the searchers were allowed as much time to formulate the Boolean queries as they felt necessary. Despite this, these queries proved less effective.

Locke et al. [72] have previously discussed Turtle's [116] criticisms of Boolean systems for, mainly, the lack of relevance ranking and result sets that are larger than what a user would be prepared to browse. This was such, as Maxwell and Schafer [79] and Turtle had noted [115, 116], that the problem with Boolean retrieval is "that the larger the collection searched on, the greater the difficulty in achieving" higher precision [72]. Schweighofer and Geist [103], in considering the need for query expansion, shared this view. As Locke et al. [72] summarized, they noted that the effectiveness of Boolean queries might not be poor in the legal domain because "lawyers as domain experts will have knowledge of synonyms, without which effectiveness may suffer ... [but] domain knowledge has its limits, and one cannot reasonably know all other possible choices for a word" [72]. This problem is common to both Boolean queries and best match retrieval. However, there is no empirical evidence to suggest this.

This brings about questions of whether explicit methods (query suggestion) or implicit methods (query expansion) can address this potential problem. We discuss these methods for case law retrieval later. Scheweighofer and Geist [103] argued that usage of term frequencies in the legal domain is not as helpful as other domains because "no redundancy exists in legal norms, but a lot of information is irrelevant in case law. Relevant texts parts may consist only of a short paragraph or even only of a single sentence in a very long legal document." Kumar [65] discussed problems posed by the language used in legal documents. But, while not specific to case law, the problems are pertinent. Given the particular legal sublanguage, Kumar states that the statistical characteristics of words may be different from those in other more general corpora, with key terms appearing few times, and that there is no factorisation of the context in which such words may

appear. Moens [83] stated that legal language is not a pure sublanguage. Rather, it consists of a diverse vocabulary usually employing domain specific concepts. Further, it is unique in so far as legal language is typically composed of "exceptionally long sentences and in crucial subclauses" [83].

In the context of web search, Carpineto et al. [21] noted in 2009 that the average query length was 2.3 words, and that this average had not increased in the 10 years prior. As a result of the shortness of queries employed by information-retrieval (IR) users, the vocabulary problem has become "even more serious", and that the size of data makes "polysemy more severe". More recent query logs have shown queries to typically be of a similar length (an average of 2.4 words) [130]. While web queries may be poor, there is no qualitative analysis to report on the effectiveness of queries in case law search. The only literature specific to case law retrieval is a comment by Shankar and Buddarapu [105], in a study of query intent classification, that queries commonly used in a commercial case law search system (LexisNexis) are very short. This leads to a poor ability to specify a user's information need. But it must be borne of mind that query length is task specific. In this regard, we note that as more question answering systems are included in commercial search systems, average query lengths may increase. This will be a result of more queries expressing information needs as questions.

Finally, the length of the result set is not a problem unique to Boolean retrieval. As previously stated, the task of case law retrieval is viewed as a recall orientated task. But it is not a total recall task. Accordingly, in favouring recall over precision, with lawyers likely unwilling to wade through all results returned, ranking and explainability of results returned by a system are key. Early Boolean retrieval models ranked documents by date, and ranked retrieval by other means was not, at this time, implemented in commercial systems. The advantage to such a system was that it was easy to explain why a document is retrieved by a query, whereas it is much more difficult to explain ranking models despite the superior effectiveness that they offered. Explainability of a search system is important for reproducibility. But, as can be contrasted with, for example, systematic reviews where there is a strict protocol for searching, or technology assisted review where there are discussions regarding query formulation, this task is interactive and reproducibility is not strictly necessary. Reproducibility is even less so important as regards the question of a lawyer's searching being competent or negligent given search history logs. The more important question is, in our view, one of user satisfaction as to why a query returns certain results. One more point may be made of reproducibility. Near constant adding of cases and updating the treatment of cases may mean that results are not reproducible.

5.2 Conceptual search, ontologies, case-based retrieval and argument retrieval

Much research has focused on ontological or conceptual methods for case law retrieval. Most of this research focuses, however, on the creation of systems for practical applications rather than the evaluation of a system's effectiveness, or on systems that match cases to a set of factual circumstances or a seed case. It is not our intention to summarise all of the research in this regard.

The problem with simple full text matching, and therefore the need for a deeper understanding of the text being searched on is said to come from, as Matthijssen [77] put it, the "conceptual gap". There is in essence, a difficulty in translating an information need into a query because it is difficult to know what information solves a legal problem prior to finding it [7]. As Carvalho [22] noted, in legal texts, concepts are often used in different ways when compared to common language. Matthijssen stated that queries are often too unspecific; the particular "information need is lost in the query formulation process". One such way to overcome this problem is enhancing browsing features. However this falls outside the scope of our survey, given our focus on ad-hoc query search. This problem can be thought of as similar to that of vocabulary mismatch, where a user may not express their information need in the same manner as that represented in the documents searched on [82]. Such a problem is common in Boolean retrieval [36]. However, given the particular legal sublanguage of case law, where concepts are expressed in similar terms, vocabulary mismatch may only pose a problem when first approaching a search task. Once a user has found one relevant document or piece of information, interactively reformatting a query to use the language of found information may alleviate this vocabulary mismatch problem.

Klein et al. [59] described ontological based methods for retrieving similar cases to a seed case. This is not in the context of ad-hoc search, but rather so as to advise parties to litigation of similar cases based on their factual situation. They index documents by removing stop-words, reducing all verbs to a particular form, and mapping documents onto a conceptual structure using an ontology. They then retrieve documents using thesaurus-based statistical retrieval to identify relevant words. From this, each document is given a "concept fingerprint", being the list of concepts identified in a document as well as scores denoting relevance of that document to a particular concept. Represented as a vector, queries are then matched to documents in a vector space. They do not achieve great effectiveness with manual or automatically created "fingerprints". Gifford [42] discussed a new commercial legal case law retrieval system designed for finding arguments. This is slightly distinct from the task of finding case law as the primary goal of the system is, as Gifford put it "for

discovering and presenting legal arguments rather than entire appellate cases." The system used hierarchies with an ability to search over any part of the hierarchy.

Matthijssen [78] detailed a task-based index structure for retrieval. The index is structured so that terms correspond to tasks, being "a group of activities and procedural steps that are directed towards a common goal". This, Matthijssen argued, simplified the gap between a lawyer expressing a particular information need and task, however it increased the gap between indexing documents and the task. To address the same problem but from the perspective of laypersons, Uijjtenbroek et al. [118] and Laarschot et al. [119] described a system for assisting the general public with case law retrieval. They used an ontology to map terms that a layperson would use to an ontology used for indexing decisions according to their legal concepts. This is one way to overcome another main problem with ontological methods, being, retrieval is dependant on expert domain knowledge [36]. Another related way to overcome this problem is, more relevant to ad-hoc query search, query expansion with ontologies.

A key issue with the use of ontologies for ad-hoc case law retrieval is that, as Maxwell and Schafer [79] recognised, with the amount of case law decided, manually annotating cases according to ontologies is not feasible. This means reliance must be placed on automatic methods for determining a case's ontological classification. Such a problem was noted as early as 2001 [83]. For this reason, focus on other methods for full-text retrieval is necessary. In somewhat of an attempt to combat this, and of relevance to ad-hoc query search, Rissland and Daniels [93] detailed the combination of case-based reasoning (CBR) systems to supplement IR systems. As they recognised, CBR systems are deeply structured but contain very few documents, whereas IR systems contain many documents but represent them at a shallow level (text only). Their system used a list of output cases from a CBR analysis as input to the INQUERY IR system [20], to provide relevance feedback on this set in order to create a query for searching over the whole collection of case law; the "use of relevance feedback, in effect, tells the IR component that the cases found through the CBR analysis are highly relevant and that the IR engine should retrieve more like them." Two limitations are evident. First, this system is limited to taking a seed case as input and does not address ad-hoc query based retrieval. Second, necessarily finding cases in the initial CBR step will be limited to the scope of the CBR collection. Such a problem is also recognised by El Jalali [36]. More recent work detailed by Conrad and Al-Kofahi [23] discusses large scale argumentation mining tools for jury verdicts, employed at Westlaw (scenario based search). Their work enables lawyers to view and group case outcomes from 500,000 cases according to a number of features. Accordingly, the

problems identified by Moens [83] and Rissland and Daniels [93] may pose less of a problem as more and more machine learning based approaches are adopted.

5.3 Question Answering

Many works have considered question answering for legal information retrieval tasks. These consider case law retrieval in that commercial systems return snippets of cases to users. More recent works by Westlaw detail many of their question answering systems [31, 80]. As is the trend with these works, and information retrieval generally, they employ machine learning methods and leverage Westlaw's vast user query logs to train the systems [90]. WestSearch PLUS [80] is the latest iteration of this. It provides a natural language interface to over 22 million single sentence summaries of case law, and high-confidence candidate answers are returned along with ad-hoc results returned from Westlaw's search system. Similarly, Bennett et al. [8] demonstrated a system employed at LexisNexis that automatically identifies parts of documents that are extracted and treated as answers to be presented to users. This system suggests potential questions to users in a query box via autocompletion and returns an answer card above ad-hoc search results. As these are commercial systems, little is detailed by way of evaluation.

There is also a large body of work that considers retrieval of civil code articles based on a sample question from the Japanese Bar exam in the COLIEE tasks [35, 55], as well as multilingual retrieval over European Union legislative materials, in the ResPubliQA challenge [88]. These works are outside the scope of this article. However, their approaches are nonetheless informative. Learning to rank and deep learning models trained using various features are commonplace. The investigation of these methods is nonetheless applicable to ranking of ad-hoc case law retrieval as commercial case law retrieval systems employ learning to rank methods [75]. The task also accounts for paragraph retrieval, which is something that has not been considered in the literature but passage or sentence level retrieval is suggested as being important to users of commercial legal search engines [73].

5.4 Query expansion

The goal of query expansion is to introduce terms in an effort to address vocabulary-mismatch problems; a goal similar to that of conceptual search methods. In Table 6, we summarise research on query expansion in the considered domain. Outside the domain and more generally, Carpineto

et al. [21] provided a survey of automatic query expansion for IR. There has been little work exploring query expansion for case law retrieval.

Custis and Al-Kofahi [29, 30] are the only works exploring query expansion not using ontologies. These works were conducted internally at Westlaw on their commercial systems and using the benefit of their extensive user logs and data. It is pertinent to note the use of query expansion in many major commercial legal search systems [126]. Such use appears common place. Earlier work of Custis and Al-Kofahi [29] automatically degraded queries by removing terms, introducing term mismatch, to better evaluate the effectiveness of their expansion methods. They compared these methods to BM25 and Query Likelihood baselines. They do not detail their expansion methods other than to say one is BM25 with pseudo-relevance feedback, and the other is a "language modeling based retrieval engine that utilizes a subject-appropriate external corpus (i.e., legal or news) as a knowledge source". The reference to Berger and Lafferty [9] vis. translation probabilities suggests it follows the same approach as their later work [30]. Custis and Al-Kofahi's empirical evaluation found that implicit expansion in their proprietary system greatly helped effectiveness. However, they found BM25 with pseudo-relevance feedback was less effective than the BM25 baseline. Over larger collections that did not contain case law, BM25 with pseudo-relevance feedback typically outperformed their proprietary method when more term mismatch was introduced at deeper measures (Recall@1000, MAP) compared to lower measures (Precision@10).

Custis and Al-Kofahi [30] do not use explicit expansion, instead relying on the retrieval model used, being the probabilistic translation language model proposed by Berger and Lafferty [9]. They consider three methods to compute the translation probabilities: (i) an estimate using co-occurence probabilities of words within a window of words surrounding a term; (ii) a PRF based measure that limits the translation component of Berger and Lafferty's model to the top 50 terms from the top 20 documents retrieved by a search; and (iii) log analysis with 39 million click through events to estimate probabilities. Their analysis involved removing query terms from a query and comparing relevant documents. They found that, over 20,000 cases and 335 queries, query expansion significantly improved both MAP and recall at 1000. The results show that simply estimating the translation probabilities using surrounding words achieves good results, even when compared to estimating probabilities using query logs. It would be interesting to see what effect however, is had at lower depths, i.e. P@20 or P@50 given the authors' view that case law retrieval is interactive and lawyers are more likely to stop searching after 20 documents and reformulate their query [115].

Work	Summary	Evaluation	Findings
[29]	Degraded query by removing query terms and evaluation of implicit query expansion proprietary concept Westlaw search system and BM25 with pseudo-relevance feedback	Evaluated over internal collection of 11,000 case law documents and 44 queries with total relevance assessments	Query expansion led to significantly less decrease in effectiveness of degraded queries, however BM25 with feedback was less effective than standalone BM25
[30]	Comparison of Berger and Lafferty's translation language model on Westlaw search system	Evaluated on internal collection of 20,000 documents and 335 queries with annotations by lawyers	Query expansion improved effectiveness compared to keyword queries
[103]	Practical implementation knowledge base query expansion system for public use	No quantitative evaluation	Query expansion improves quality of search results
[100]	Practical implementation using ontology based expansion for retrieval of legal audio-video media	No quantitative evaluation	Query expansion was promising
[15]	Evaluation of ontologies automatically expanded from a seed ontology for finding relevant sentences	Evaluated over 2 collections of a total of 1,303 sentences	No comparison to a baseline method without query expansion

Table 6. Summary of works evaluating query expansion.

We also note that recent works in IR have proposed alternative methods for estimating $P(q|D)$, the effect of which in this domain may be interesting to investigate [52, 53, 133].

It is worth mentioning that use of query logs for case law retrieval is becoming more frequent. Their use can be seen in training legal question-answering systems [31, 90] as well as in recommending cases to users based on similar user profiles [73] (another tool outside of ad-hoc querying that is of great use to lawyers once they have found a relevant case for a given information need). Learning-to-rank based approaches adopted at Westlaw [75] might lead one to hypothesise that query logs are also applied to ad-hoc ranking for case law retrieval too.

Ontological framework based methods represent the bulk of, and the remaining extent of, the literature as regards query expansion. The study by Breuker et al. [18] is one of the earliest suggestions of using ontology based query expansion for case law retrieval. They suggested expansion with both terms that are superclasses, and also terms that are subclasses, of terms specified by a user. Similarly, Sartori et al. [100] and Schweighofer and Geist [103] have both proposed query expansion in the legal domain using ontological frameworks. Scheweighofer and Geist [103], do

not report quantitative results, other than to say that the improvement is as expected, "quite good". Likewise, Sartori [100] does not provide any detail of improvements in performance. Other research into query expansion for case law retrieval also lacks quantitative evaluation [17, 108, 131]. Saravanan and Gavindran [98] considered manual query expansion as well as suggestion through an interface that suggests terms based on an ontological framework. While they report greater performance in terms of precision and recall, they do not consider automated methods for expansion (users expansions are selected by users). Similarly, Tantisripreecha and Soonthornphisaj [110] considered query expansion by means of a domain ontology to improve result diversity. They create a set of queries where each contains an ontological concept, based on concepts related to the original query concept. They then issued these queries and rank each query individually, combining all results at the end to compile a ranked result list. They noted increases in diversity of results. Again however, their testing on two collections, each of less than 1000 cases, and just over 10 topics may mean that their results do not generalise.

Boonchom and Soonthornphisaj [15] described a manual ontology based expansion system, where by traversing an ontology more terms are suggested to the user. The system also weighted ontological concepts. They reported precision over a small collection of 1300 sentences and thus their empirical findings may not generalise. This collection is split into two distinct topics, however the size of each is not known. On one collection, the weighted ontology obtained a precision of 0.89, whereas the non-weighted ontology obtained a precision of 0.72. Results on the other collection were lower for both weighted and unweighted approaches. Unfortunately their work does not compare the ontologies to a baseline method for retrieval of the sentences.

Finally and relatedly, Savelka et al [101] attempted query expansion when searching for sentences in case law that discuss statutory provisions. They did this by two means: including words from the provision itself; and, taking similar words from word embeddings learned over the collection of cases to those words found in the query. But their evaluation found that expansion by selecting similar terms did not increase retrieval effectiveness.

5.5 Query reduction

Query reduction consists of selecting a subset of terms from a query to construct a more effective query [72]. This is typically in the context of more verbose queries. As with other facets of case law retrieval, little work has examined query reduction. We summarise these works in Table 7: the

Work	Summary	Evaluation	Findings
[72]	Reduction methods based on several term scoring methods	Evaluated over collection of 60,000 cases and 100 queries	Query reduction improved effectivenss of retrieval system
[71]	Reduction methods based on several term scoring methods	Evaluated over collection of 3 million cases and 12 queries	Query reduction improved effectivenss of retrieval system
[70]	Reduction methods based on several term scoring methods used to find cited cases	Evaluated on collection of cases where cited cases were anonymised	Reduction is promising method to identify cited cases

Table 7. Summary of works evaluating query reduction.

Work	Summary	Evaluation	Findings
[61]	Diversification using various methods	Evaluated on collection of 60,000 cases with automatically generated relevance assessments	Search result diversification methods outperform text summarisation methods and `tf-idf` baseline

Table 8. Summary of works evaluating search result diversification.

only works that have examined the application of keyword extraction or query reduction methods are those of Locke et al. [72] and Locke and Zuccon [70, 71].

Locke et al. [72] considered several measures for term scoring for automatic reduction of text from cases and compared these to the effectiveness of manually generated queries. They took a similar approach to that described in the work of Koopman et al. [63] with regard to generating queries for a range of proportions (Koopman et al. considered query reduction in the medical domain). To score terms for inclusion in a query, they evaluated IDF, parsimonious language model and Kullback-Leibler informativeness. Similar approaches were adopted in Locke and Zuccon [71], as well as in Locke and Zuccon [70], but this was not in the context of ad-hoc query retrieval, rather in seeking to find cited cases within a case given the text surrounding a reference to a case.

Research in the domain has only considered unigram based statistical methods for selection of appropriate terms when automatically generating queries. Logically therefore, methods that focus on the multi-word terms (n-grams) or syntactic phrases may be an appropriate next step for consideration of automatic query reduction methods.

5.6 Search Result Diversification

We summarise the works involving diversification in Table 8. Diversification requires trading off finding relevant documents for a diverse set of documents for a given query [96]. It is important to

note that this may not be an appropriate task given the suggested nature of case law retrieval as a recall orientated task, but one concerned with precision. If the task is one of finding relevant seed documents the time taken to complete a given task is of key importance. The ranking of results may differ for a diversified set, and this will lead to an effect on the time taken for a searcher to find all relevant decisions and to complete a search task. However, this is not something that can be easily quantified, nor has research attempted to do so. But, if case law retrieval is a total recall task, diversification, which will only affect the ranking of documents and not the retrieved set, will merely only affect user satisfaction with a given system.

Koniaris et al. [61] stated that "it is extremely difficult to search for relevant [cases] by using Boolean queries", and that diverse results are "intuitively" more informative than homogenous result sets that contain results with similar features. They concluded that diversification methods "demonstrate notable improvements in terms of enriching search results with other hidden aspects of the legal query space", and after empirically evaluating the methods according to a-nDCG (alpha-normalised discounted cumulative gain), ERR-IA (expected reciprocal rank intent aware) and S-recall (subtopic recall), web search diversification methods were the best methods evaluated. Koniaris et al. [61] sought to return a set of 30 diversified results from an original set of 100 after an original ranking. They considered these results at cutoffs of 5, 10, 20 and 30 for varying levels of interpolation between relevance and diversity, following a ranking of results produced using cosine similarity and log-based TF-IDF methods. While various methods of diversification are considered, they evaluated their methods over a collection of 3890 cases, for which they used automatically obtained relevance assessments from an LDA topic model trained over the highest scored TF-IDF documents for a query. The results may not be generalisable, and may be different had manual relevance assessments been obtained.

5.7 Use of citation networks

Citations are an important part of a legal decision. Judges cite for a number of reasons. In most instances, there is an obligation to give reasons for a decision. For a judge to show that their decision conforms with established and binding principle, they must explain their conclusion. In so doing, they will state the law, by reference to a particular decision, or to another source of law. They will, broadly speaking, say why their decision follows a particular law, or does not. A citation by a judge of a previous decision may express a range of acceptance of the cited decision, from

Work	Summary	Evaluation	Findings
[111]	Earliest proposal for use of citations to rank case law by representing a case as a vector of citations	Manual comparison of seed to highest ranked cases	Citation vectors provide a means for finding similar cases
[126]	Comparison of citation networks in Dutch case law	Manual evaluation of citation types in case law	Citations are likely a useful feature for ranking case law
[39, 40]	Use of citations as feature in vector-space model in commercial system	Not considered	Not considered
[61]	Use of citations to diversify case law results	Cranfield style evaluation on automatically created collection	Citations provide effective diversification of case law search results

Table 9. Summary of works evaluating use of citation networks for case law retrieval.

following the decision, to distinguishing (saying that it does not apply to the factual situation), to doubting its application, or rejecting it.

Citation analysis in case law has been frequently studied. For instance, Agnoloni et al. [3], Fowler [37], Neale [84], Koniaris et al. [62], Geist [38] and van Opijnen [120] present large studies of citation networks in case law. With regard to case law retrieval, citation indexes have existed for some time [132].[15] Works by persons at Westlaw have said that citations are used for ranking case law [24, 116], as well as their use in other search systems [75], but, unsurprisingly, there has been little publicised application of network analysis methods towards improving effectiveness of case law retrieval systems. We summarise these works in Table 9.

The earliest proposal of the use of citations in case law retrieval is by Tapper [111]. First as a means of tracing relevant cases [111]. Several commercial systems provide tools for doing this, by visualising a decision's citation network.[16] Later, Tapper proposed citations as a means for ranking by representing cases as a vector of their citations, as opposed to a vector of its terms [112]. Despite: (i) the creation of visual tools for mapping citation networks; (ii) suggestions that citations are a useful ranking feature [126]; and (iii) findings that legal citation networks exhibit similarities as compared to other networks [38] that have been used to rank results of search systems, citation networks have only been applied to ranking documents for case law retrieval in two instances. This can be contrasted with the employment of such measures in web result ranking [81, 86].

[15] Two commercial services exist: LexisNexis Shepard and Westlaw Keycite. As do other freely available systems such as LawCite: www.austlii.edu.au/lawcite.
[16] See www.ravellaw.com and www.jade.io.

The FLEXICON system combined citations as part of the ranking process in a vector space model for searching over a very small collection of case law (1000 documents) [39, 40]. As a commercial system, the exact way in which citations were used was not detailed.

Koniaris et al. [61] considered an application of DivRank in seeking to diversify search results. They found it, and other web-based diversification methods were effective.

While not in the context of reranking results for ad-hoc queries, Winkels et al. [128], used network analysis to suggest decisions that are relevant to a particular legislative provision. This approach is similar to Tapper's early approach for finding similar cases to a seed case using citation vectors [112], where documents were represented as a series of citation vectors rather than term vectors. This has one notable drawback: use of citation vectors only to rank cases rather than as a feature means that cases that are not cited, or do not cite other cases will not be ranked. Such an outcome is undesirable and may happen where a case considers a new, uncommon, point of law, or the case is newly decided and has not been cited by any other cases.

As Winkels et al. [129] noted, many of the most cited cases (in a network of Dutch cases) deal with procedural issues rather than substantive issues. Substantive laws are those which identify and define rights, whereas procedural laws deal only with regulating court proceedings: *McKain v RW Miller & Co (SA) Pty Ltd* (1991) 174 CLR 1.

This focus of cases on procedural issues is illustrative of one of the problems with the use of a network to rank documents. Panagis et al. [87] recognised the different circumstances in which a decision may be cited and that simple network analysis of citations falls short in this regard by equally treating citations; a citation may be made without much application of any principle or it may be made for a number of other reasons, from approving, following or dissenting from (expressing disagreement with), a decision. Unlike webs page where it may be safe to assume that the page deals only with one topic, case law frequently involves multiple topics. For instance, a case may decide any number of questions, each unrelated to the other. In this respect, topicality of the citation, or the manner in which the cited case is treated are interesting areas that have not yet been explored. Any consideration of citations for ranking should consider both the topicality of the citation and the treatment of the cited case.

While not for ad-hoc search, Zhang and Koppaka [132] from LexisNexis, proposed a system that analysed the citation network of a case based on the topic for which a case was cited. This system allowed users to follow links based on the legal topic. The system involved discerning

the reasons for which a case was cited through looking at the citation's surrounding sentences. Similarly, commercial systems offer tools for tracing citations at a paragraph level.[17]

In terms of such tools being incorporated into ad-hoc retrieval, Panagis et al. [87] explored citations at a paragraph level rather than at a document level, in combination with text analysis to determine non-explicit references. But, a limitation of their work is that they did not consider ranking. Raghav et al. [92] ranked case law based on a combination of two measures. The first measure split a seed judgment into a set of paragraphs and calculated BM25 similarity between the paragraphs of all other judgments. The second measure used was bibliographic coupling to calculate similarity between two judgments. Bibliographic coupling is the number of citations to another document that two documents share [54]. This ranking was based on an existing seed judgment rather than a query, however. Given the above comments regarding topicality at a paragraph level, the use of bibliographic coupling is odd, given that this will calculate the similarity between judgments overall based on the similarity of their citations. More recently, more complex methods for identifying similarities between cases have been proposed. These take into account not just a computed similarity between the citations of two cases, but also the similarity between references to legislation [11].

5.8 Deeper understanding of legal text and AI applications to case law retrieval

Applying AI methods to legal problems has been a topic of substantial interest in the past 20 or so years. The International Conference for Artificial Intelligence and Law shows a depth of work in this area. We do not wish to repeat much of the work that has occurred in this domain, as it is on other legal tasks and outside of the scope of this survey. However, for useful surveys of the works in ICAIL see Conrad and Zeleznikow [25, 26].

Shankar and Buddarapu [106] detail character level machine translation models for query reformulation, word-embedding based query-intent understanding [19], as well deep ensemble learning query understanding techniques [105] used at LexisNexis. Query reformulation and understanding all form important parts of the ad-hoc retrieval process, ensuring queries are routed to a correct vertical search engine and any entities in a query are correctly identified. As Shankar and Buddarapu recognise, this represents a step away from traditional search and towards predictive coding. Other applications of AI to law by Westlaw and LexisNexis include automatic detection of overruling of cases and parsing cases for use in litigation analytics tools [31].

[17] Jade Barnet offers such a tool: see jade.io/j/?ht=FAQ+-+Focus+Matches&t=help.

Work	Summary	Evaluation	Findings
[43]	Annotated portions of a case for use refining space of document text searched on	Non-Cranfield style evaluation over collection of 188 cases	Poor classification of sentences detrimentally affected retrieval performance
[85]	Annotated sentences to train classifier to then search for sentences relevant to a query	Non-Cranfield style evaluation	Annotating sentences is effective means for finding factually similar sentences

Table 10. Summary of works evaluating semantic roles of text.

Semantic search for ad-hoc case law retrieval has begun to be investigated. Relatedly, see the above discussion on query expansion based on translation probabilities. Given the move towards predictive coding, this is an area that warrants further investigation. Only one work has considered ranking case law using similarities between embeddings for sentences or paragraphs of a case [99]. Sarsa and Hyvonen [99] described a prototype system that ranked Finnish case law given a seed document using, amongst other methods, Doc2Vec [68]. This is an analog to systems provided by commercial search providers, that combine numerous features including recommending based on similar user profiles [73]. Savelka et al [101] used similarities between embeddings to find sentences in case law that are relevant to statutory provisions. Landthaler et al. [66] used word-embeddings to find similar legal provisions in documents and in statutes. Word embedding similarities have also seen use in other legal tasks such as e-Discovery [122]. No work has specifically considered leveraging semantic similarities for ranking for ad-hoc case law retrieval.

Closely related, two works have considered the impact of classification of text according to its rhetorical role, as a means of retrieving relevant decisions [43, 85]. We summarise these works in Table 10. We note that similar applications in the context of literature retrieval exist [16, 102].

Semantic roles of text refers to what role the text plays within the document: i.e. is it conclusive or factual. In case law this will be whether the text is factual or legal. As the doctrine of precedent requires that like circumstances be considered in a similar fashion, being able to search a collection solely for similar factual situations may prove useful to lawyers. Similarly, lawyers as domain experts are often required to apply reasoning from other legal areas. As a result, the ability of a lawyer to search for law that is logically similar but not factually relevant may also prove useful. As an example, the meaning of words are often key to interpreting a piece of statute – if the statute prohibits something that *may* give rise to another thing, the meaning of *may* could be determined by reference to legal decisions considering the word. And, the ability to search for legal statements

that explain the meaning of the word would therefore prove useful. The intuition is therefore that searching on a confined part of the whole collection may lead to both precision and recall.

The only study that has considered such an approach while ranking results is that of Grabmair et al. [43], who evaluated retrieval in the context of searching through a collection of 188 annotated vaccine related adverse reaction legal cases. They considered annotation of rhetorical roles for the purpose of ranking. They retrieved all sentences relevant to a query and then ranked documents based on the number of relevant sentences a document contains. Again to be noted is the small collection on which evaluation is undertaken, a common problem with case law retrieval evaluations.

The only other work is that of Nejadghole et al. [85], who implemented a semantic search system that searched only on facts of cases. They did so by training a classifier on a small number of annotated cases, and a word embedding model to automatically annotate a collection. They manually annotated 150 decisions out of a collection of 46,000 decisions, for a total of 12,220 sentences. Their annotations classified 46% of the sentences in these judgments as fact. From their manual annotations, they trained a binary classifier to automatically annotate the remaining decisions in their collection. In order to evaluate retrieval performance, they created 15 queries, combined with 5 sentences for each query to compare whether they were in fact relevant or not relevant. They ordered results according to similarity to the query. They reported an *MAP* of 0.78. While this research reports an interesting application of methods applied outside of legal IR to the retrieval of case law, it does not do so in a way that uses such a method for ranking in ad-hoc querying. By way of comparison, there have been several applications of rhetorical roles to ranking in the task of retrieval of systematic reviews [16, 102].

There are obvious difficulties in retrieval based on this approach. Moens [83] identified that low density training examples pose a problem for text categorisation. This is ultimately the problem that Grabmair et al [43] faced, given that ranking ignoring a sentence's classification resulted in better performance. In related work in another professional search field, Boudin [16] stated that the performance of a weighted field approach may not work well because of difficulty arriving "at a consistent tagging of [the] elements." Such a difficulty is nonetheless present in annotating case law. But, there would appear to be large amounts of work in the automatic summarisation field that could be leveraged [43, 46, 97, 107]. Again, we note that this is extremely similar to tools such as Westlaw's scenario search, the difference being the use of natural language querying as opposed

to searching through filtering different fields [23]. As we noted above, the use of machine learning techniques appears to have done away with much of the problem identified by Moens [83].

6 OPEN ISSUES, CHALLENGES AND DIRECTIONS

Improving the effectiveness of ad-hoc case law search still faces fundamental challenges. Despite the size of the legal research industry and the legal industry that relies on it, little has been published on ranking for ad-hoc retrieval in the last 20 years. Further, there is no standard collection for evaluating methods, and this results in the difficulty in comparing any potential methods. The main focus of work in legal IR is on the semantic understanding of case law text, or legal text in general, or technology-assisted review. This work on semantic understanding is promising, but tends to be performed on a much smaller scale compared to typical IR evaluations: hundreds of documents instead of hundreds of thousands or millions of documents, with the exceptions of works by commercial legal search providers. Similarly, evaluations are not typical Cranfield style evaluations, but as we identified above, this is only a problem in an academic settings, and not commercially. Both of these problems mean that the conclusions reached by any evaluations may not hold true on larger collections or in real world situations.

Intrinsic characteristics of case law documents make problems very difficult in this domain. As discussed in Section 3, the extreme length, number of diverse topics in a single document, complex language and structure of the texts' pose unique difficulties not often present in other domains. Likewise, debate about whether the task is one requiring total recall or, whether as in the authors' view, it is one requiring high recall but still requiring acceptable levels of precision introduces different problems. If the system is one requiring total recall replicability, explainability and user trust are paramount [94]. Ranking is less important. If the task is precision orientated, finding relevant documents within a small amount of interactions and therefore ranking is more important.

As to appropriate future directions of research, outside establishing baselines and creation of a robust test collection on which to evaluate the effectiveness of methods, several areas warrant exploration. Firstly, the impact of domain specific features on ranking in case law retrieval: citations, which are prominent in case law, topicality, treatment and the relationship between a citing court and cited court, as well as the time between the cited decision are all matters not considered in the literature. Any such exploration of these features has not empirically evaluated their use for case law retrieval. Secondly, semantic representations of text are a typical focus point in many areas

of research. This has focused on ontological representations of text, or QA systems, all of which have been on a small scale, with the exception of recent works by commercial search providers. Consideration of semantic search, or segmenting texts according to the role a subsection plays within an overall document all represent interesting directions for research.

7 CONCLUSION

Case law is judge made law that forms one of the key areas of law in common law legal systems. Case law retrieval systems are a key tool for lawyers to be able to effectively and cost efficiently carry out their work. Despite large commercial systems existing for a number of years, little research compares the effectiveness of methods for ad-hoc query retrieval of case law. This survey has summarised the extent of literature in this regard.

In light of the existing literature, several key observations may be made. In the last 30 years, the literature has reported a turn towards natural language queries. This is an incident of Boolean retrieval becoming ever less common [13, 89] and lack of teaching in legal education will only further this. However, Boolean searches, while less frequent than natural language searches, are still an important aspect of case law search [106]. More and more research involves the application of machine learning techniques to both information retrieval and to legal tasks generally. Much of which has focused on conceptual search, and question answering. Question answering systems are taking the forefront in commercial legal search systems. Little research has been published as to ranking for ad-hoc retrieval, as opposed to finding similar judgments to seed cases. This is not surprising for several reasons: the wealth of research in other legal search areas such as e-Discovery and other assistive legal search tools; and further, given the size and competitiveness of the legal search market, commercial operators may be unwilling to disclose the factors behind any success. Query expansion is adopted by many commercial search systems but there exists little published research empirically evaluating the efficacy of such approaches. The lack of standardised testing over large collections is one of the biggest problems in ad-hoc case law retrieval and creating a collection will be a time consuming approach, that most likely must be done by lawyers. But this is not a problem for commercial search providers, who can leverage the use of vast query logs [73, 80].

Undoubtedly, the future of case law retrieval will be natural language based as Boolean queries are taught less in legal education [13], and question answering systems take the forefront in commercial search platforms. But, for effective research to take place, outside of commercial ventures,

large scale collections are required, so too are baselines of other state-of-the-art retrieval methods from other IR domains.

REFERENCES

[1] 2018. Ravel Judge Analytics. ravellaw.zendesk.com/hc/en-us/article_attachments/115014213887/Better_Motions_with_Ravel_s_Judge_and_Court_Analytics.pdf. (2018). Accessed: 2018-05-14.

[2] 2019. Senior Software Engineer - Search, Recommendation and Data infrastructure. (2019). www.efinancialcareers.com.au/jobs-USA-NY-New_York-Senior_Software_Engineer_-_Search_Recommendation_and_Data_infrastructure.id07682125

[3] Tommaso Agnoloni. 2014. Network Analysis of Italian Constitutional Case Law. In *Semantic Processing of Legal Texts*, Vol. 91. 24.

[4] Jason R Baron, Jack G Conrad, Amanda Jones, David D Lewis, and Douglas W Oard. 2015. Report of the DESI Workshop. In *ICAIL 2015 DESI VI Workshop on Using Machine Learning and Other Advanced Techniques to Address Legal Problems in E-Discovery and Information Governance.*

[5] Jason R Baron, David D Lewis, and Douglas W Oard. 2006. TREC 2006 Legal Track Overview. In *Proceedings of the 15th Text REtrieval Conference.*

[6] Jason R Baron, Douglas W Oard, David D Lewis, and Paul Thompson. 2007. In *ICAIL 2007 DESI Workshop: Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings.*

[7] Nicholas J Belkin, Robert N Oddy, and Helen M Brooks. 1982. ASK for Information Retrieval: Part I. Background and Theory. *Journal of Documentation* 38, 2 (1982), 61–71.

[8] Zachary Bennett, Tony Russell-Rose, and Kate Farmer. 2017. A Scalable Approach to Legal Question Answering. In *Proceedings of the 16th International Conference on Articial Intelligence and Law.* 269–270.

[9] Adam Berger and John Lafferty. 1999. Information Retrieval as Statistical Translation. In *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 222–229.

[10] Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019. Overview of the FIRE 2019 AILA Track: Artificial Intelligence for Legal Assistance.. In *FIRE (Working Notes).* 1–12.

[11] Paheli Bhattacharya, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. 2020. Hier-SPCNet: A Legal Statute Hierarchy-based Heterogeneous Network for Computing Legal Case Document Similarity. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1657–1660.

[12] Jon Bing. 2010. Let there be LITE: a Brief History of Legal Information Retrieval. *European Journal of Law and Technology* 1, 1 (2010).

[13] Barbara Bintliff and Duncan Alford. 2013. *Teaching Legal Research.* Taylor & Francis.

[14] 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.

[15] Vi-sit Boonchom and Nuanwan Soonthornphisaj. 2010. Thai Succession and Family Law Ontology Building Using Ant Colony Algorithm. In *New Frontiers in Artificial Intelligence.* Springer Berlin Heidelberg, 19–32.

[16] Florian Boudin, Jian-Yun Nie, and Martin Dawes. 2010. Clinical Information Retrieval Using Document and PICO Structure. In *Human Language Technologies: The 2010 Conference of the North American Chapter of the Association for Computational Linguistics.* 822–830.

[17] Joost Breuker, Abdullatif Elhag, Emil Petkov, and Radboud Winkels. 2002. Ontologies for Legal Information Serving and Knowledge Management. In *Proceedings of the 15th Annual Conference on Legal Knowledge and Information Systems.* 1–10.

[18] Joost Breuker, André Valente, and Radboud Winkels. 2004. Legal Ontologies in Knowledge Engineering and Information Management. *Artificial Intelligence and Law* 12, 4 (2004), 241–277.

[19] Venkata Nagaraju Buddarapu and Arunprasath Shankar. 2019. Data Shift in Legal AI Systems.. In *Proceedings of the 17th International Conference on Artificial Intelligence and Law.*

[20] James P Callan, W Bruce Croft, and Stephen M Harding. 1992. The INQUERY Retrieval System. In *Database and Expert Systems Applications.* Springer, 78–83.

[21] Claudio Carpineto and Giovanni Romano. 2012. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Computing Surveys (CSUR)* 44, 1 (2012), 1–50.

[22] Danilo S. Carvalho, Minh-Tien Nguyen, Chien-Xuan Tran, and Minh-Le Nguyen. 2017. *Lexical-Morphological Modeling for Legal Text Analysis.* Springer International Publishing, Cham, 295–311.

[23] Jack G Conrad and Khalid Al-Kofahi. 2017. Scenario Analytics: Analyzing Jury Verdicts to Evaluate Legal Case Outcomes. In *Proceedings of the 16th International Conference on Artificial Intelligence and Law.* 29–37.

[24] Jack G Conrad and Qiang Lu. 2013. VoxPopuLII. (Mar 2013). blog.law.cornell.edu/voxpop/2013/03/28/next-generation-legal-search-its-already-here/

[25] Jack G Conrad and John Zeleznikow. 2013. The Significance of Evaluation in AI and law. In *Proceedings of the 14th International Conference on Artificial Intelligence and Law.* 186–191.

[26] Jack G Conrad and John Zeleznikow. 2015. The Role of Evaluation in AI and Law: an Examination of its Different Forms in the AI and Law Journal. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law.* 181–186.

[27] Gordon V Cormack, Maura R Grossman, Bruce Hedin, and Douglas W Oard. 2010. Overview of the TREC 2010 Legal Track. In *Proceedings of the 19th Text REtrieval Conference*, Vol. 1.

[28] W Bruce Croft, Donald Metzler, and Trevor Strohman. 2015. *Search engines: Information Retrieval in Practice.* Addison-Wesley Reading.

[29] Tonya Custis and Khalid Al-Kofahi. 2007. A new approach for evaluating query expansion: query-document term mismatch. In *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 575–582.

[30] Tonya Custis and Khalid Al-Kofahi. 2008. Investigating External Corpus and Clickthrough Statistics for Query Expansion in the Legal Domain. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management.* ACM, 1363–1364.

[31] Tonya Custis, Frank Schilder, Thomas Vacek, Gayle McElvain, and Hector Martinez Alonso. 2019. Westlaw Edge AI Features Demo: KeyCite Overruling Risk, Litigation Analytics, and WestSearch Plus. In *Proceedings of the 17th International Conference on Artificial Intelligence and Law.* 256–257.

[32] Daniel P Dabney. 1986. The Curse of Thamus: An Analysis of Full-text Legal Document Retrieval. *Law Library Journal* 78 (1986), 5.

[33] Daniel P Dabney. 1993. *Statistical Modeling of Relevance Judgments for Probabilistic Retrieval of American Case Law.* Ph.D. Dissertation. University of California, Berkeley.

[34] Chief Justice Paul de Jersey. 2013. Practice Direction Number 16 of 2013 (Supreme Court of Queenlsand). (2013).

[35] Phong-Khac Do, Huy-Tien Nguyen, Chien-Xuan Tran, Minh-Tien Nguyen, and Minh-Le Nguyen. 2017. Legal Question Answering using Ranking SVM and Deep Convolutional Neural Network. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2016 Workshops, LENLS, HAT-MASH, AI-Biz, JURISIN and SKL.*

[36] Soufiane El Jelali, Elisabetta Fersini, and Enza Messina. 2015. Legal Retrieval as Support to eMediation: Matching Disputant's Case and Court Decisions. *Artificial Intelligence and Law* 23, 1 (2015), 1–22.

[37] James H Fowler, Timothy R Johnson, James F Spriggs, Sangick Jeon, and Paul J Wahlbeck. 2007. Network Analysis and the Law: Measuring the Legal Importance of Precedents at the US Supreme Court. *Political Analysis* (2007), 324–346.

[38] Anton Geist. 2009. *Using Citation Analysis Techniques for Computer-Assisted Legal Research in Continental Jurisdictions.* Master's thesis. University of Edinburgh.

[39] Daphne Gelbart and JC Smith. 1990. Toward a Comprehensive Legal Information Retrieval System. In *Database and Expert Systems Applications.* Springer, 121–125.

[40] Dephne Gelbart and JC Smith. 1991. Beyond Boolean Search: FLEXICON, a Legal Tex-based Intelligent System. In *Proceedings of the 3rd International Conference on Artificial intelligence and Law.* ACM, 225–234.

[41] Kevin Gerson. 1999. Evaluating Legal Information Retrieval Systems: How do the Ranked-retrieval Methods of WESTLAW and LEXIS Measure Up? *Legal Reference Services Quarterly* 17, 4 (1999), 53–67.

[42] Matthew Gifford. 2017. LexrideLaw: an Argument Based Legal Search Engine. In *Proceedings of the 16th International Conference on Artificial Intelligence and Law.* ACM, 271–272.

[43] Matthias Grabmair, Kevin D. Ashley, Ran Chen, Preethi Sureshkumar, Chen Wang, Eric Nyberg, and Vern R. Walker. 2015. Introducing LUIMA: An Experiment in Legal Conceptual Retrieval of Vaccine Injury Decisions Using a UIMA Type System and Tools. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law.* 69–78.

[44] Graham Greenleaf, Andrew Mowbary, and David Lewis. 1988. *Australasian Computerised Legal Information Handbook.* Butterworths Australia.

[45] Maura R Grossman and Gordon V Cormack. 2010. Technology-Assisted Review in e-Discovery can be More Effective and More Efficient than Exhaustive Manual Review. *Richmond Journal of Law & Technology* 17 (2010), 1.

[46] Ben Hachey and Claire Grover. 2006. Extractive Summarisation of Legal Texts. *Artificial Intelligence and Law* 14, 4 (2006), 305–345.

[47] William G Harrington. 1984. A Brief History of Computer-Assisted Legal Research. *Law Library Journal* 77 (1984), 543.

[48] Seongwan Heo, Kihyun Hong, and Young-Yik Rhim. 2017. Legal Content Fusion for Legal Information Retrieval. In *Proceedings of the 16th International Conference on Artificial Intelligence and Law.* ACM, 277–281.

[49] Julian PT Higgins and Sally Green. 2011. *Cochrane Handbook for Systematic Reviews of Interventions.* Vol. 4. John Wiley & Sons.

[50] Peter Jackson, Khalid Al-Kofahi, Alex Tyrrell, and Arun Vachher. 2003. Information Extraction from Case Law and Retrieval of Prior Cases. *Artificial Intelligence* 150, 1-2 (2003), 239–290.

[51] Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. 2019. COLIEE-2018: Evaluation of the Competition on Legal Information Extraction and Entailment. In *New Frontiers in Artificial Intelligence.* 177–192.

[52] Maryam Karimzadehgan and ChengXiang Zhai. 2010. Estimation of Statistical Translation Models Based on Mutual Information for Ad Hoc Information Retrieval. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 323–330.

[53] Maryam Karimzadehgan and ChengXiang Zhai. 2012. Axiomatic Analysis of Translation Language Model for Information Retrieval. In *European Conference on Information Retrieval.* Springer, 268–280.

[54] Maxwell Mirton Kessler. 1963. Bibliographic Coupling Between Scientific Papers. *American Documentation* 14, 1 (1963), 10–25.

[55] Mi-Young Kim and Randy Goebel. 2017. Two-step Cascaded Textual Entailment for Legal Bar Exam Question Answering. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law.* 283–290.

[56] Mi-Young Kim, Ying Xu, and Randy Goebel. 2017. *Applying a Convolutional Neural Network to Legal Question Answering.* Springer International Publishing, Cham, 282–294.

[57] Mi-Young Kim, Ying Xu, Randy Goebel, and Ken Satoh. 2014. *Answering Yes/No Questions in Legal Bar Exams.* Springer International Publishing, Cham, 199–213.

[58] Mi-Young Kim, Ying Xu, Yao Lu, and Randy Goebel. 2016. Legal Question Answering Using Paraphrasing and Entailment Analysis. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2016 Workshops, LENLS, HAT-MASH, AI-Biz, JURISIN and SKL.*

[59] Michel CA Klein, Wouter van Steenbergen, Elisabeth M Uijttenbroek, Arno R Lodder, and Frank van Harmelen. 2006. Thesaurus-based Retrieval of Case Law. In *Proceedings of the 2006 Conference on Legal Knowledge and Information Systems.* IOS Press, Amsterdam, 61–70.

[60] Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. 2016. Multi-dimension Diversification in Legal Information Retrieval. In *International Conference on Web Information Systems Engineering.* Springer, 174–189.

[61] Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. 2017. Evaluation of Diversification Techniques for Legal Information Retrieval. *Algorithms* 10, 1 (2017), 22.

[62] Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. 2018. Network Analysis in the Legal Domain: A Complex Model for European Union Legal Sources. *Journal of Complex Networks* 6 (2018), 243–268. Issue 2.

[63] Bevan Koopman, Liam Cripwell, and Guido Zuccon. 2017. Generating clinical queries from patient narratives. In *Proceedings of the 40th international ACM SIGIR conference on Research and Development in Information Retrieval.*

[64] Yogesh Kulkarni, Rishabh Patil, and Srinivasan Shridharan. 2017. Detection of Catchphrases and Precedence in Legal Documents. In *Working notes of FIRE 2017-Forum for Information Retrieval Evaluation.* 86–89.

[65] Sushanta Kumar. 2014. *Similarity Analysis of Legal Judgements and applying Paragraph-link to Find Similar Legal Judgments.* Master's thesis. International Institute of Information Technology Hyderabad.

[66] Jörg Landthaler, Bernhard Waltl, Patrick Holl, and Florian Matthes. 2016. Extending Full Text Search for Legal Document Collections Using Word Embeddings.. In *Legal Knowledge and Information Systems: JURIX 2016: The 29th Annual Conference.*

[67] Steven A Lastres. 2013. *Rebooting Legal Research in a Digital Age.* Technical Report. LexisNexis.

[68] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *International Conference on Machine Learning.* 1188–1196.

[69] Wenhui Liao and Isabelle Moulinier. 2009. Feature Engineering on Event-Centric Surrogate Documents to Improve Search Results. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management.* 1629–1632.

[70] Daniel Locke and Guido Zuccon. 2017. Automatic Cited Decision Retrieval: Working Notes of Ielab for FIRE Legal Track Precedence Retrieval Task. In *Working notes of FIRE 2017-Forum for Information Retrieval Evaluation (CEUR Workshop Proceedings).* 80–81.

[71] Daniel Locke and Guido Zuccon. 2018. A Test Collection for Evaluating Legal Case Law Search. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 1261–1264.

[72] Daniel Locke, Guido Zuccon, and Harrisen Scells. 2017. *Automatic Query Generation from Legal Texts for Case Law Retrieval.* Springer International Publishing, 181–193.

[73] Qiang Lu and Jack G Conrad. 2012. Bringing Order to Legal Documents. In *International Conference on Knowledge Engineering and Ontology Development.*

[74] Susan Nevelow Mart. 2013. The Case for Curation: the Relevance of Digest and Citator Results in Westlaw and Lexis. *Legal Reference Services Quarterly* 32, 1-2 (2013), 13–53.

[75] Susan Nevelow Mart, Joe Breda, Ed Walters, Tito Sierra, and Khalid Al-Kofahi. 2019. Inside the Black Box of Search Algorithms. *AALL Spectrum* (2019).

[76] Dean Mason. 2006. Legal Information Retrieval Study: Lexis Professional and Westlaw UK. *Legal Information Management* 6, 04 (2006), 246–250.

[77] Luuk Matthijssen. 1995. An Intelligent Interface for Legal Databases. In *Proceedings of the 5th International Conference on Artificial Intelligence and Law.* ACM, 71–80.

[78] Luuk Matthijssen. 1998. A Task-Based Interface to Legal Databases. *Artificial Intelligence and Law* 6, 1 (1998), 81–103.

[79] Tamsin Maxwell and Burkhard Schafer. 2008. Concept and Context in Legal Information Retrieval. In *Proceedings of the 21st Conference on Legal Knowledge and Information Systems.* IOS Press, Amsterdam, 63–72.

[80] Gayle McElvain, George Sanchez, Sean Matthews, Don Teo, Filippo Pompili, and Tonya Custis. 2019. WestSearch Plus: A Non-Factoid Question-Answering System for the Legal Domain. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1361–1364.

[81] Donald Metzler. 2011. *A Feature-Centric View of Information Retrieval.* The Information Retrieval Series, Vol. 27. Springer Berlin Heidelberg.

[82] Donald Metzler, Susan Dumais, and Christopher Meek. 2007. Similarity Measures for Short Segments of Text. In *European Conference on Information Retrieval.* Springer, 16–27.

[83] Marie-Francine Moens. 2001. Innovative Techniques for Legal Text Retrieval. *Artificial Intelligence and Law* 9, 1 (2001), 29–57.

[84] Thom Neale. 2013. Citation Analysis of Canadian Case Law. *Journal of Open Access to Law* 1 (2013), 1.

[85] Isar Nejadgholi, Renaud Bougueng, and Samuel Witherspoon. 2017. A Semi-Supervised Training Method for Semantic Search of Legal Facts in Canadian Immigration Cases. *Legal Knowledge and Information Systems* (2017), 125.

[86] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report. Stanford InfoLab.

[87] Yannis Panagis, Urska Sadl, and Fabien Tarissan. 2017. Giving Every Case its (Legal) Due: The Contribution of Citation Networks and Text Similarity Techniques to Legal Studies of European Union law. In *Proceedings of the 30th International Conference on Legal Knowledge and Information Systems*, Vol. 302. IOS Press, 59–68.

[88] Anselmo Peñas, Pamela Forner, Richard Sutcliffe, Álvaro Rodrigo, Corina Forăscu, Iñaki Alegria, Danilo Giampiccolo, Nicolas Moreau, and Petya Osenova. 2010. Overview of ResPubliQA 2009: Question Answering Evaluation Over European Legislation. *Multilingual Information Access Evaluation I. Text Retrieval Experiments* (2010), 174–196.

[89] Joshua Poje. 2014. Legal Research. *American Bar Association Techreport 2014* (2014).

[90] Filippo Pompili, Jack G Conrad, and Carter Kolbeck. 2019. Exploiting Search Logs to Aid in Training and Automating Infrastructure for Question Answering in Professional Domains. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. 93–102.

[91] 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*. 34–49.

[92] K Raghav, P Krishna Reddy, and V Balakista Reddy. 2016. Analyzing the Extraction of Relevant Legal Judgments using Paragraph-level and Citation Information. *Artificial Intelligence for Justice* (2016), 30.

[93] Edwina L Rissland and Jody J Daniels. 1996. The Synergistic Application of CBR to IR. *Artificial Intelligence Review* 10, 5-6 (1996), 441–475.

[94] Tony Russell-Rose, Jon Chamberlain, and Leif Azzopardi. 2018. Information Retrieval in the Workplace: A Comparison of Professional Search Practices. *Information Processing & Management* 54, 6 (2018), 1042–1057.

[95] Mark Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends® in Information Retrieval* 4, 4 (2010), 247–375.

[96] Rodrygo LT Santos, Craig Macdonald, Iadh Ounis, et al. 2015. Search Result Diversification. *Foundations and Trends® in Information Retrieval* 9, 1 (2015), 1–90.

[97] M Saravanan, Balaraman Ravindran, and S Raman. 2006. Improving Legal Document Summarization using Graphical Models. *Frontiers in Aritifical Intelligence and Applications* 152 (2006), 51–60.

[98] M. Saravanan, B. Ravindran, and S. Raman. 2009. Improving Legal Information Retrieval using an Ontological Framework. *Artificial Intelligence and Law* 17, 2 (2009), 101–124.

[99] Sami SARSA and Eero Hyvonen. 2019. Searching Case Law Judgements by Using Other Judgements as a Query. (2019).

[100] Fabio Sartori and Matteo Palmonari. 2010. *Query Expansion for the Legal Domain: A Case Study from the JUMAS Project.* Springer Berlin Heidelberg, Berlin, 107–122.

[101] Jaromir Savelka, Huihui Xu, and Kevin D Ashley. 2019. Improving Sentence Retrieval from Case Law for Statutory Interpretation. In *Proceedings of the 17th International Conference on Artificial Intelligence and Law*. ACM, 113–122.

[102] Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Leif Azzopardi, and Shlomo Geva. 2017. Integrating the Framing of Clinical Questions via PICO into the Retrieval of Medical Literature for Systematic Reviews. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2291–2294.

[103] Erich Schweighofer and Anton Geist. 2007. Legal Query Expansion Using Ontologies and Relevance Feedback. In *Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques*. 149–160.

[104] Walid Shalaby and Wlodek Zadrozny. 2019. Patent retrieval: a literature review. *Knowledge and Information Systems* (2019), 1–30.

[105] Arunprasath Shankar and Venkata Nagaraju Buddarapu. 2018. Deep Ensemble Learning for Legal Query Understanding. In *International Workshop on Legal Data Analytics and Mining*.

[106] Arunprasath Shankar and Venkata Nagaraju Buddarapu. 2019. Legal Query Reformulation using Deep Learning. In *Proceedings of the 3rd Workshop on Automated Semantic Analysis of Information in Legal Text*.

[107] Olga Shulayeva, Advaith Siddharthan, and Adam Wyner. 2016. Recognizing Cited Facts and Principles in Legal Judgements. *Artificial Intelligence for Justice* (2016), 52.

[108] Jayasudha Subburaj, C Soundarya Veni, and R Nandhini Amirtha. 2016. Legal Text Retrieval–Semantic Web Approach. *International Journal of Advanced Research Trends in Engineering and Technology* 3(11) (2016), 86–91.

[109] John I Tait. 2014. An Introduction to Professional Search. In *Professional Search in the Modern World*. Springer, 1–5.

[110] Tanapon Tantisripreecha and Nuanwan Soonthornphisaj. 2009. Query Expansion Algorithm for Supreme Court Sentences Retrieval Using Ontology. In *Proceedings of the 48th Kasetsart Annual Conference*. 43–50.

[111] Colin Tapper. 1974. Legal Information Retrieval by Computer: Applications and Implications. *McGill Law Journal* 20 (1974), 26.

[112] Colin Tapper. 1981. The Use of Citation Vectors for Legal Information Retrieval. *Journal of Law and Information Science* 1 (1981), 131.

[113] Luiyang Tian, Hui Ning, Leilei Kong, Zongyuan Han, Ruiming Xiao, and Haoliang Qi. 2017. HLJIT2017@IRLed-FIRE2017: Information Retrieval From Legal Documents. In *Working notes of FIRE 2017-Forum for Information Retrieval EvaluationRab*. 82–85.

[114] Howard Turtle. 1991. *Inference Networks for Document Retrieval.* Ph.D. Dissertation. University of Massachussets.

[115] Howard Turtle. 1994. Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance. In *Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 212–220.

[116] Howard Turtle. 1995. Text Retrieval in the Legal World. *Artificial Intelligence and Law* 3, 1 (1995), 5–54.

[117] Howard Turtle and W Bruce Croft. 1991. Evaluation of an Inference Network-Based Retrieval Model. *ACM Transactions on Information Systems* 9, 3 (1991), 187–222.

[118] Elisabeth Uijttenbroek, Arno Lodder, Michel Klein, Gwen Wildeboer, Wouter Van Steenbergen, Rory Sie, Paul Huygen, and Frank van Harmelen. 2008. *Retrieval of Case Law to Provide Layman with Information about Liability: Preliminary Results of the BEST-Project.* Springer Berlin Heidelberg, 291–311.

[119] Ronny van Laarschot, Wouter Van Steenbergen, Heiner Stuckenschmidt, Arno R. Lodder, and Frank van Harmelen. 2005. The Legal Concepts and the Layman's Terms - Bridging the Gap through Ontology-Based Reasoning about Liability. In *Proceedings of the 18th Conference on Legal Knowledge and Information Systems.* 115–125.

[120] Marc van Opijnen. 2012. Citation Analysis and Beyond: in Search of Indicators Measuring Case Law Importance. In *Proceedings of the 18th Conference on Legal Knowledge and Information Systems*, Vol. 250. 95–104.

[121] Marc van Opijnen and Cristiana Santos. 2016. On the Concept of Relevance in Legal Information Retrieval. *Artificial Intelligence for Justice* (2016), 78.

[122] Ngoc Phuoc An Vo, Caroline Privault, and Fabien Guillot. 2017. Experimenting Word Embeddings in Assisting Legal Review. In *Proceedings of the 16th International Conference on Articial Intelligence and Law.* 189–198.

[123] Ellen M Voorhees, Donna K Harman, et al. 2005. *TREC: Experiment and Evaluation in Information Retrieval.* Vol. 63.

[124] Pengfei Wang, Yu Fan, Shuzi Niu, Ze Yang, Yongfeng Zhang, and Jiafeng Guo. 2019. Hierarchical Matching Network for Crime Classification. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 325–334.

[125] Pengfei Wang, Ze Yang, Shuzi Niu, Yongfeng Zhang, Lei Zhang, and ShaoZhang Niu. 2018. Modeling Dynamic Pairwise Attention for Crime Classification over Legal Articles. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval.* 485–494.

[126] Gineke Wiggers and Suzan Verberne. 2019. Citation Metrics for Legal Information Retrieval Systems. *Proceedings of the 8th Workshop on Bibliometric-enhanced Information Retrieval* (2019), 39–50.

[127] Robert A Wilson. 1962. Computer Retrieval of Case Law. *Southwestern Law Journal* 16 (1962), 409.

[128] Radboud Winkels, Alexander Boer, Bart Vredebregt, and Alexander van SOMEREN. 2014. Towards a Legal Recommender System. In *Proceedings of the 27th International Conference on Legal Knowledge and Information Systems.* 169–179.

[129] Radboud Winkels, Jelle de Ruyter, and Henryk Kroese. 2011. Determining Authority of Dutch Case Law. *Legal Knowledge and Information Systems* 235 (2011), 103–112.

[130] Lee Xiong, Chuan Hu, Chenyan Xiong, Daniel Campos, and Arnold Overwijk. 2019. Open Domain Web Keyphrase Extraction Beyond Language Modeling. (2019), 5178–5187.

[131] Soraya Zaidi, Mohamed Tayeb Laskri, and K Bechkoum. 2005. A Cross-language Information Retrieval Based on an Arabic Ontology in the Legal Domain. In *Proceedings of the International Conference on Signal-Image Technology and Internet-Based Systems.* 86–91.

[132] Paul Zhang and Lavanya Koppaka. 2007. Semantics-based Legal Citation Network. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law.* 123–130.

[133] Guido Zuccon, Bevan Koopman, Peter Bruza, and Leif Azzopardi. 2015. Integrating and Evaluating Neural Word Embeddings in Information Retrieval. In *Proceedings of the 20th Australasian Document Computing Symposium.* 1–8.