Artificial Intelligence and Law manuscript No.
(will be inserted by the editor)

Evaluation of Information Retrieval for E-Discovery

Douglas W. Oard · Jason R. Baron · Bruce Hedin · David D. Lewis · Stephen Tomlinson

Received: December 24, 2009 / Accepted: April 16, 2010

Abstract The effectiveness of information retrieval technology in electronic discovery (e-discovery) has become the subject of judicial rulings and practitioner controversy. The scale and nature of e-discovery tasks, however, has pushed traditional information retrieval evaluation approaches to their limits. This paper reviews the legal and operational context of e-discovery and the approaches to evaluating search technology that have evolved in the research community. It then describes a multi-year effort carried out as part of the Text Retrieval Conference to develop evaluation methods for responsive review tasks in e-discovery. This work has led to new approaches to measuring effectiveness in both batch and interactive frameworks, large data sets, and some surprising results for the recall and precision of Boolean and statistical information retrieval methods. The paper concludes by offering some thoughts about future research in both the legal and technical communities toward the goal of reliable, effective use of information retrieval in e-discovery.

Keywords E-Discovery · Information Retrieval · Interactive Search · Evaluation

The first three sections of this article draw upon material in the introductory sections of two papers presented at events associated with the 11th and 12th International Conferences on Artificial Intelligence and Law (ICAIL) (Baron and Thompson, 2007; Zhao et al, 2009) as well as material first published in (Baron, 2008), with permission.

Douglas W. Oard
College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 USA, E-mail: oard@umd.edu

Jason R. Baron
Office of the General Counsel, National Archives and Records Administration, College Park, MD 20740 USA, E-mail: jason.baron@nara.gov

Bruce Hedin
H5, 71 Stevenson St., San Francisco, CA 94105 USA, E-mail: bhedin@h5.com

David D. Lewis
David D. Lewis Consulting, 1341 W. Fullerton Ave., #251, Chicago, IL 60614 USA, E-mail: ailj10paper@DavidDLewis.com

Stephen Tomlinson
Open Text Corporation, Ottawa, Ontario, Canada, E-mail: stomlins@opentext.com

1 Introduction

The legal profession and its corporate and other institutional clients are increasingly confronting a new reality: massive and growing amounts of electronically stored information (ESI) required to be retained under both records laws[1] and in light of pending litigation (The Sedona Conference, 2007a). The phenomenon known as "e-discovery" (i.e., the requirement that documents and information in electronic form stored in corporate systems be produced as evidence in litigation) is sweeping the legal profession in the U.S., and is a potential international force for change in the ways institutions of all stripes manage and preserve their proprietary information. Not a week now goes by without a new decision issued by a court which turns on the failure to find and/or preserve important evidentiary information in electronic form—sometimes resulting in sanctions to lawyers and their clients.[2] In turn, a spotlight has now formed on how lawyers decide to meet their obligations in various e-discovery contexts. A major aspect of this involves how they go about directing "searches" for relevant electronic evidence in response to discovery requests or due to some other external demand for information coming from a court or an inquiring party with standing to make such a request. The "how" is increasingly important, given spiraling legal costs. For example, a Forrester report issued in 2006 estimated that the secondary market for information technology solutions in the legal technology sector will grow just in the U.S. to $4.8 billion by 2011.[3]

The potential magnitude of the search problem is highlighted by past research indicating that lawyers greatly overestimate their true rate of recall in civil discovery (i.e., how well their searches for responsive documents have uncovered all relevant evidence, or at least all potential "smoking guns"). A seminal study evaluating the success of text retrieval methods in a legal setting was conducted by Blair and Maron in 1985 (Blair and Maron, 1985). That study established a gap between the perception on the part of lawyers that using their specific queries they would retrieve on the order of 75% of the relevant evidence to be found in a collection of 40,000 documents gathered for litigation purposes, whereas the researchers were able to show that only about 20% of relevant documents had in fact been found.

The unprecedented size, scale, and complexity of electronically stored data now potentially subject to routine capture in litigation presents Information Retrieval (IR) researchers with a series of important challenges to overcome, not the least of which is a fundamental question as to how best to model the real world. At least two of the major research efforts on legal applications aimed at evaluating the efficacy of the search task—the Blair and Maron study and the Text Retrieval Conference (TREC) Legal Track—required manual assessments of the responsiveness of approximately 3×10^4 documents (drawn from a much larger population of documents of 7×10^6 for the TREC Legal Track). These past efforts have utilized certain designs and evaluation criteria that may or may not prove to be optimal for future research projects involving data sets

[1] See, e.g., Sarbanes-Oxley Act, Title 18 of the U.S. Code, Section 1519 (U.S. securities industry requirement to preserve email for 7 years); National Archives and Records Administration regulations, 36 Code of Federal Regulations Part 1236.22 (all email that is considered to fall within the definition of "federal records" under Title 44 of the U.S. Code, Section 3301, must be archived in either paper or electronic systems).

[2] See, e.g., Qualcomm v. Broadcom, 539 F. Supp. 2d 1214 (S.D. Cal 2007), rev'd 2010 WL 1336937 (S.D. Cal. Apr. 2, 2010); In re Fannie Mae Litigation, 552 F.3d 814 (D.C. Cir. 2009).

[3] http://www.forrester.com/Research/Document/Excerpt/0,7211,40619,00.html

with perhaps orders of magnitude of more responsive and non-responsive documents. It is now well understood that as data sets get larger, high-precision searches generally become somewhat easier, but "indeterminacy multiplies making it increasingly difficult to conduct successful specific or exhaustive searches" (Blair, 2006). Thus, faced with a full spectrum of candidate search methods, we may legitimately ask: are the evaluation measures in present use adequate to explore the range of research questions we need to consider? If not, what new developments are needed?

We begin our investigation in Section 2 by examining how the context of law, and the practice of law, affects the nature of text search performed in e-discovery. In Section 3 we examine changing judicial views of what has become known as "keyword search" in the legal profession, ending with a cautionary note on the differences between the ways similar technical vocabulary have been used by practitioners of e-discovery and by IR researchers. Section 4 reviews the history of work on evaluation of IR, culminating in the well known TREC conferences. Section 5 describes our efforts to bring TREC-style evaluation to bear on e-discovery problems through the TREC Legal Track. We discuss the Interactive, Ad Hoc, and Relevance Feedback tasks, as well as the TREC tobacco and email test collections. We look beyond TREC in Section 6 to discuss large-scale situated studies, going beyond search to modeling the full e-discovery process, and what a research path towards certifying "process quality" for e-discovery might look like. We conclude in Section 7 with some final thoughts.

2 The Legal Context

As an initial step in thinking about how to structure IR research for the purpose of advancing our knowledge about improving the efficacy of legal searches in a real world context, three types of relevant factors potentially serve to inform the discussion: (i) the size and heterogeneity of data sets made subject to discovery in current litigation; (ii) what the nature of the legal search task is perceived by lawyers to be; and (iii) how the search function is actually performed by real lawyers and agents acting on their behalf in concrete situations. A fourth factor, namely, the degree to which the legal profession can be expected to absorb new ways in which to do business, or to tolerate alternative methodologies, is optimistically assumed, but not further considered here. Note that for present purposes, we primarily focus on the experience of lawyers in civil litigation within the U.S., although the principles discussed would be expected to have broader application.

2.1 Size and Heterogeneity Issues

An unquantified but substantial percentage of current litigation is conducted by parties holding vast quantities of evidence in the form of ESI. Directly as the result of the unparalleled volume and nature of such newly arising forms of evidence, Congress and the Supreme Court approved changes to the Federal Rules of Civil Procedure, in effect as of December 1, 2006, which inter alia added "ESI" as a new legal term of art, to supplement traditional forms of discovery wherever they may have previously pertained or applied to mere "documents." As just one example of this phenomenon, 32 million email records from the White House were made subject to discovery in U.S. v. Philip Morris, the racketeering case filed in 1999 by the Justice Department against several tobacco corporations. Out of the subset represented by 18 million presidential record emails, using a search method with hand-built queries combining search terms using Boolean, proximity and truncation operators, the government uncovered 200,000 potentially responsive electronic mail (email) messages, many with attachments. These in turn were made subject to further manual review, on a one-by-one basis, to determine responsiveness to the litigation as well as status as privileged documents. The review effort required 25 individuals working over a six month period (Baron, 2005). Apart from this one case, it appears that in a number of litigation contexts over 10^9 electronic objects have been preserved for possible access, as part of ongoing discovery (Jensen, 2000).[4] Accordingly, the volume of material presented in many current cases precludes any serious attempt being made to solely rely on manual means of review for relevance. Thus, "in many settings involving electronically stored information, reliance solely on a manual review process for the purpose of finding responsive documents may be infeasible or unwarranted. In such cases, the use of automated search methods should be viewed as reasonable, valuable, and even necessary."[5] However, greater reliance on automated methods will in turn raise questions of their accuracy, efficacy, and completeness.

In addition to exponential increases in volume, the collections themselves are rapidly evolving. The past decade has seen not only an explosion in email traffic, but also the growth of dynamic structured databases and unstructured content of all kinds, including universes of data found in Web 2.0 applications, both in corporate intranets as well as out in the Internet "cloud," and in all formats ranging from audio to video to virtual reality (e.g., Second Life). Electronic storage devices have similarly evolved rapidly, thus making the search problem one needing to encompass evidence stored on all forms of current and legacy media, hard drives, network servers, backup tapes, and portable devices of all kinds. In turn, the universe of data made available to the typical end user performing functions on today's corporate desktop computers constitutes a prime target in civil litigation. Thus, the subject of search is a worthy candidate for academic research evaluating the efficacy of search methods.

2.2 Nature of the Legal Search Task

For the most part discovery is conducted by means of inquiring on an open-ended basis into selected topics of relevance to a particular case, including through depositions, interrogatories, and requests to produce documents (including now ESI). Although exceptional situations arise where lawyers are focused on retrieving a known small set of one or more particular documents, in the vast majority of cases the lawyers' inquiry in discovery is intended to be broadly worded, to capture "any and all" (or certainly as many as possible) relevant pieces of evidence to the case at hand. Thus, the "ad hoc" nature of the lawyer's search task. In turn, "relevance" is defined broadly under the law: if any fragment of text in a document has bearing on a contested issue, the document as a whole is found to be relevant and should presumptively be turned over to opposing counsel, absent assertion of privilege.

[4] See also Report of Anton R. Valukas, Examiner, In re Lehman Brothers Holdings Inc. (U.S. Bankruptcy Ct. S.D.N.Y. March 11, 2010), vol. 7, Appx. 5 (350 billion pages subjected to dozens of Boolean searches), available at http://lehmanreport.jenner.com/

[5] See Practice Point 1 in (The Sedona Conference, 2007b) (referred to herein as the "Sedona Search Commentary").

2.3 How Legal Searches Are Actually Performed

The state of practice as it generally exists today consists of lawyers, in response to broadly worded discovery demands, directing key custodians and their IT staff counterparts to search for relevant information using a set of keywords dreamed up unilaterally, with limited use made of Boolean, proximity, or other operators. It may strike some as incredible that until only very recently lawyers in some of the most complex litigations failed to employ any particularly advanced strategies for developing search protocols as an aid in conducting searches—and that this situation still continues to this day as the status quo reality, rather than being the exceptional case. This entire area of the law is now changing, however, with the advent of the new federal rules of civil procedure, and the emergence of a body of case law questioning the ad hoc, unexplained, and/or unilaterally deployed use of single keywords as search terms.

As a starting proposition, it is well understood, at least in the U.S., that "broad discovery is a cornerstone of the litigation process contemplated by the Federal Rules of Civil Procedure."[6] In other words, "fishing expeditions" seeking the broadest amount of evidence have been encouraged at all levels of the legal profession, as an "engine" of the discovery process. How lawyers go about propounding and responding to discovery didn't materially change between the 1930s and the 1990s (and for some increasingly isolated practitioners, has never changed). Under well-known U.S. rules of civil procedure governing the so-called "discovery" phase of civil litigation prior to trial, lawyers constructed what are known as "interrogatories" and "document requests," as two staples of the art of discovery practice. Document requests (i.e., requests that the other side produce documents relevant to a named topic) were propounded with the expectation that the producing party would perform a reasonably diligent search for records found in corporate hardcopy repositories—including file room areas and work stations. Although the rules require lawyers to certify that they have performed a "complete" good faith response to discovery requests, a "perfect" search has never been required.[7] A party has always, however, had the right to challenge the adequacy of a given search for relevant documents, if they have reason to believe based on documents produced (if any) that an opposing party failed to account for all known sources of relevant evidence. The latter could include a failure to check with all reasonable custodians of documents, including key players known to be material witnesses in the litigation.

During these seven decades, "motion practice" over document requests usually has consisted of sequences of interactions akin to chess moves: one party crafting requests in the broadest conceivable way; the producing party reflexively opposing such language as overbroad, vague, ambiguous, and not leading to the discovery of relevant evidence; the requesting party filing a motion to compel a ruling from the court on the ambiguity inherent in the requests, with a demand that the adversary "do something" to respond to the original queries; and the court finally stepping in at some later stage in the process, to assist in the crafting of narrower or more precise inquiries and to require production under a fixed deadline. All of this litigation gamesmanship was routinely carried out prior to any "meeting of the minds" by the parties to attempt to resolve the scope of production amicably, and prior to any production whatsoever of actual

[6] Zubulake v. UBS Warburg LLC, 217 F.R.D. 309, 311 (2003); see generally, The Sedona Conference, The Sedona Principles (2d ed. 2007).

[7] See Pension Committee of the University of Montreal Pension Plan et al. v Banc of America Securities LLC, et al., 2010 WL 184312, *1 (S.D.N.Y. Jan. 15, 2010) ("Courts cannot and do not expect that any party can reach a standard of perfection.").

relevant documents by either side in a case. Objectively speaking, there was some measure of rationality in not proceeding to undertake responses to discovery where the relevant evidence at issue “merely” consisted of hardcopy documents in traditional file cabinets and boxes. Any search meant hours of intensive labor manually performing file and document level review for relevant evidence, plus a second pass to determine potentially “privileged” documents out of those documents segregated as responsive to particular requests.

This general model for civil discovery carried even into the era of office automation; however, the growth of networks and the Internet, resulting in exponential increases in ESI, has changed the legal terrain considerably. Lawyers familiar with structured collections of cases and legislation, as created by Lexis and Westlaw, became readily adept at using basic Boolean, and occasionally ranked retrieval, searches on free text and/or controlled vocabulary terms to find relevant precedents. To the same end, lawyers also increasingly were able to utilize simple search strategies to tackle the task of finding relevant documents in poorly structured corporate document collections.

3 The Jurisprudence of “Keyword” Search

As IR researchers have long known, and recent legal scholarship has recognized, text retrieval systems suffer from a variety of limitations, given the inherent ambiguity of language, the well characterized limitations in the ability of people to formulate effective queries, and further complexities introduced by preprocessing (e.g., optical character recognition for scanned documents).[8] Nor did the mere utilization of automated means for conducting searches change the basic procedural equation between legal adversaries throughout discovery (since there is an inherent asymmetry in the positions of parties with respect to the state of their knowledge of what relevant documents exist—with one side flying completely “blind”).[9]

Only a few years ago, the idea that there would be a jurisprudence devoted to analyzing the strengths and limitations of text retrieval would be unheard of, for at least two reasons. First, until recent times, discovery practice, even in the largest and most complex cases in the U.S., consisted entirely of paper productions of documents, sometimes in admittedly massive quantities. For example, just short of ten million documents in hardcopy form have been amassed in a repository, pursuant to a Master Settlement Agreement, between a variety of State Attorneys General and the tobacco industry.[10]

[8] See (The Sedona Conference, 2007b), at 202.

[9] There is, at virtually all times, an admitted asymmetry of knowledge as between the requesting party (who does not own and therefore does not know what is in the target data collection), and the receiving or responding party (who does own the collection and thus in theory could know its contents). For an exploration into ethical questions encountered when the existence of documents is not reached by a given keyword search method, see (Baron, 2009).

[10] For example, see, People of the State of California v. Philip Morris, et al., Case No. J.C.C.P. 4041 (Sup. Ct. Cal.) (December 9, 1998 consent decree incorporating terms of Master Settlement Agreement or “MSA”). These documents have for the most part been digitized using Optical Character Recognition (OCR) technology and are available online on various Web sites. See the Legacy Tobacco Collection, available at http://legacy.library.ucsf.edu/. The OCR portions of the MSA collection have been used in conjunction with the TREC Legal Track.

Second, with the advent of proprietary computerized databases of case law, represented most prominently by Westlaw and Lexis, lawyers have become well versed in conducting searches to find relevant case precedent for use in legal pleadings and briefs. The beauty of text retrieval in this context is that no lawyer wishes or needs to read more than a handful of cases as an aid in stating a legal position in writing, except in the rare instance where more exhaustive searches are necessary. Further, databases of case law contain few enough and valuable enough items that vendors can apply rich controlled vocabulary indexing, in addition to the crucial content indicators provided by citations among cases.

In contrast, the limitations of simple search techniques become more apparent as the task changes from finding case precedent to finding "all" relevant evidence related to the discovery topic at hand. This has become especially clear with the exponential growth of databases and data stores of all kinds, especially with respect to email (Paul and Baron, 2007).

3.1 Keywords, Concepts, and the Courts

Discussions of the challenges of search in the e-discovery context have focused on the perceived weaknesses of "keyword" search, a point of terminology to which we will return in Section 3.2. The Sedona Conference's commentary on search and IR methods presents the issues this way:

> Keyword searches[11] work best when the legal inquiry is focused on finding particular documents and when the use of language is relatively predictable. For example, keyword searches work well to find all documents that mention a specific individual or date, regardless of context. However, . . . the experience of many litigators is that simple keyword searching alone is inadequate in at least some discovery contexts. This is because simple keyword searches end up being both over- and under-inclusive in light of the inherent malleability and ambiguity of spoken and written English (as well as all other languages).[12]

The Sedona Search Commentary describes how keywords have the potential to miss documents that fail to contain the word either because other terms with the same meaning have been used, or due to common or inadvertently misspelled instances of the keyword term. The Commentary then goes on at length to describe some of the technologies that have been proposed for dealing with weaknesses of simple term matching, including Boolean operators, fuzzy logic, statistical techniques, and taxonomies and ontologies.[13] The Commentary makes the point that lawyers "are beginning to feel more comfortable" using these forms of alternative search tools, based on anecdotal evidence from a small (but increasing) number of companies and law firms.[14]

While the limitations of simple term matching have been well-known to the IR community for decades, U.S. jurisprudence has only recently begun to grapple with the problems inherent in the task of conducting a reasonable search for all of the

[11] As used by e-discovery practitioners, "keyword search" most often refers to the use of single query terms to identify the set of all documents containing that term as part of a pre-processing step to identify documents that merit manual review.

[12] (The Sedona Conference, 2007b) at 201. See also (Paul and Baron, 2007).

[13] Id. at 202–03; 217 (Appendix describing alternative search methods at greater length).

[14] Id. at 202–03.

needles in what turn out to be very large e-haystacks. In a short interval of time, some courts have gone from extolling the power of keyword searching to questioning its efficacy (at least as articulated by counsel). Compare the case of In re Lorazepam & Clorazepate Antitrust Litigation, 300 F. Supp. 2d 43, 46 (D.D.C. 2004):

> "[t]he glory of electronic information is not merely that it saves space but that it permits the computer to search for words or 'strings' of text in seconds,"

to U.S. v. O'Keefe, 537 F. Supp. 2d 14, 24 (D.D.C. 2008):

> "Whether search terms of 'keywords' will yield the information sought is a complicated question involving the interplay, at least, of the sciences of computer technology, statistics, and linguistics. . . . Given this complexity, for lawyers and judges to dare opine that a certain search term or terms would be more likely to produce information than the terms that were used is truly to go where angels fear to tread."

Until mid-2007, the overarching approach taken by a number of courts in this area has been to define the reasonableness of the search conducted by a party solely by the number of keyword terms being requested and their relevance to the subject at hand. Thus, in the case of In re Lorazepam, the district court endorsed the employment of a number of search terms as a reasonable means of narrowing the production for relevant ESI.[15] In another case, as few as four keyword search terms were found to be sufficient.[16] In certain decisions, the court ordered a producing party (usually the defendant) to conduct searches using the keyword terms provided by plaintiff.[17] More recently, judges that have taken a more activist approach have attempted to force parties to cooperate on reaching an agreement for a reasonable search protocol, including the use of certain search terms.[18]

On June 1, 2007, U.S. Magistrate Judge John Facciola issued an opinion in the case of Disability Rights Council of Greater Washington v. Metropolitan Transit Authority[19] in which for the first time in published case law a judge suggested that parties contemplate the use of an alternative to merely reaching a set of keywords by consensus. The dispute in question involved disabled individuals and an advocacy group bringing an action against a local transit authority alleging that inadequacies in paratransit services amounted to disability discrimination. The plaintiffs moved to compel the production of electronic documents residing on backup tapes in the defendants' possession. After engaging in a routine balancing analysis of the considerations set out

[15] In re Lorazepam & Clorazepate Antitrust Litigation, 300 F. Supp. 2d 43 (D.D.C. 2004).

[16] J.C. Associates v. Fidelity & Guaranty Ins. Co., 2006 WL 1445173 (D.D.C. 2006).

[17] For example, see Medtronic Sofamor Danck, Inc. v. Michelson, 229 F.R.D. 550 (W.D. Tenn. 2003); Treppel v. Biovail, 233 F.R.D. 363, 368–69 (S.D.N.Y. 2006) (court describes plaintiff's refusal to cooperate with defendant in the latter's suggestion to enter into a stipulation defining the keyword search terms to be used as a "missed opportunity" and goes on to require that certain terms be used); see also Alexander v. FBI, 194 F.R.D. 316 (D.D.C. 2000) (court places limitations on the scope of plaintiffs' proposed keywords in a case involving White House email).

[18] In addition to cases discussed infra, see, e.g., Dunkin Donuts Franchised Restaurants, Inc. v. Grand Central Donuts, Inc, 2009 WL 175038 (E.D.N.Y. June 19, 2009) (parties directed to meet and confer on developing a workable search protocol); ClearOne Communications, Inc. v. Chiang, 2008 WL 920336 (D. Utah April 1, 2008) (court adjudicates dispute over conjunctive versus disjunctive Boolean operators).

[19] 242 F.R.D. 139 (D.D.C. 2007)

in Rule 26(a) of the Federal Rules of Civil Procedure, the court ordered that some form of restoration of the backup tapes be ordered to recover relevant documents. It is at this juncture that the opinion broke new ground: the magistrate judge expressly required that counsel meet and confer and prepare for his signature a "stipulated protocol" as to how the search of the backup tapes would be conducted, and pointed out "I expect the protocol to speak to at least the following concerns," including both "How will the backup tapes be restored?", and

> "Once restored, how will they be searched to reduce the electronically stored information to information that is potentially relevant? In this context, I bring to the parties' attention recent scholarship that argues that concept searching, is more efficient and more likely to produce the most comprehensive results."[20]

Following this decision, Judge Facciola, writing in U.S. v. O'Keefe,[21] chose to include a discussion of the use of search protocols. The O'Keefe case involved the defendant being indicted on the charge that as a State Department employee living in Canada, he received gifts and other benefits from his co-defendant, in return for expediting visa requests for his co-defendant's company employees. The district court judge in the case had previously required that the government "conduct a thorough and complete search of both its hardcopy and electronic files in a good faith effort to uncover all responsive information in its possession custody or control."[22] This in turn entailed a search of paper documents and electronic files, including for emails, that "were prepared or received by any consular officers" at various named posts in Canada and Mexico "that reflect either policy or decisions in specific cases with respect to expediting visa applications."[23]

The defendants insisted that the government search both active servers and certain designated backup tapes. The government conducted a fairly well-documented search, as described in a declaration placed on file with the court, in which 19 specific named individuals were identified as being within the scope of the search, along with certain identified existing repositories by name and the files of at least one former member of staff. The declarant went on to describe the search string used as follows:

> "early or expedite* or appointment or early & interview or expedite* & interview."[24]

Upon review of the results, only those documents "clearly about wholly unrelated matters" were removed, for example, "emails about staff members' early departures or dentist appointments." Nevertheless, the defendants objected that the search terms used were inadequate. This led the magistrate judge to state that on the record before him, he was not in a position to decide whether the search was reasonable or adequate, and that given the complexity of the issues he did not wish "to go where angels fear to tread." The court went on to note, citing to the use of "expert" testimony under Federal Rule of Evidence 702:

[20] Id. at 148 (citing to (Paul and Baron, 2007), supra).

[21] 537 F. Supp. 2d 14, 24 (D.D.C. 2008).

[22] Id. at 16 (quoting U.S. v. O'Keefe, 2007 WL 1239204, at *3 (D.D.C. April 27, 2007)) (internal quotations omitted).

[23] 537 F. Supp. 2d at 16.

[24] Based only on what is known from the opinion, it is admittedly somewhat difficult to parse the syntax used in this search string. One is left to surmise that the ambiguity present on the face of the search protocol may have contributed to the court finding the matter of adjudicating a proper search string to be too difficult a task.

> "This topic is clearly beyond the ken of a layman and requires that any such conclusion be based on evidence that, for example, meets the criteria of Rule 702 of the Federal Rules of Evidence. Accordingly, if defendants are going to contend that the search terms used by the government were insufficient, they will have to specifically so contend in a motion to compel and their contention must be based on evidence that meets the requirements of Rule 702 of the Federal Rules of Evidence."[25]

Whether it is the view of the magistrate judge that expert opinion testimony must be introduced in all cases on the subject of the reasonableness of the search method or protocol employed immediately generated discussion in subsequent case law and commentary.[26]

In 2008, Magistrate Judge Paul Grimm further substantially contributed to the development of a jurisprudence of IR through issuance of a comprehensive opinion on the subject of privilege review in Victor Stanley, Inc. v. Creative Pipe, Inc.[27] At issue was whether the manner in which privileged documents were selected from a larger universe of relevant evidence was sufficient to protect a party from waiver of attorney-client privilege, where 165 privileged documents were provided to the opposing counsel as the result of a keyword search. At the outset, Judge Grimm reported that "he ordered the parties' computer forensic experts to meet and confer in an effort to identify a joint protocol to search and retrieve relevant ESI" in response to the plaintiff's document requests. The protocol "contained detailed search and information retrieval instructions, including nearly five pages of keyword/phrase search terms."[28]

The defendants' counsel subsequently informed the court that they would be conducting a separate review to filter privileged documents from the larger [universe] of 4.9 gigabytes of text-searchable files and 33.7 gigabytes of non-searchable files. In doing so, they claimed to use seventy keywords to distinguish privileged from non-privileged documents; however, Judge Grimm, applying a form of heightened scrutiny to the assertions of counsel, found that their representations fell short of being sufficient for purposes of explaining why mistakes took place in the production of the documents and in so doing, avoiding waiver of the privilege. In the court's words:

> "[T]he Defendants are regrettably vague in their description of the seventy keywords used for the text-searchable ESI privilege review, how they were developed, how the search was conducted, and what quality controls were employed to assess their reliability and accuracy. . . . [N]othing is known from the affidavits provided to the court regarding their [the parties' and counsel's] qualifications for designing a search and information retrieval strategy that could be expected to produce an effective and reliable privilege review. . . .
> [W]hile it is universally acknowledged that keyword searches are useful tools for search and retrieval of ESI, all keyword searches are not created equal; and there is a growing body of literature that highlights the risks associated with

[25] 537 F. Supp. 2d at 24.

[26] Equity Analytics v. Lundin, 248 F.R.D. 331 (D.D.C. 2008) (stating that in O'Keefe "I recently commented that lawyers express as facts what are actually highly debatable propositions as to efficacy of various methods used to search electronically stored information," and requiring an expert to describe scope of proposed search); see also discussion of Victor Stanley, Inc. v. Creative Pipe, Inc., infra.

[27] 250 F.R.D. 251 (D. Md. 2008).

[28] Id. at 254.

> conducting an unreliable or inadequate keyword search of relying exclusively on such searches for privilege review."[29]

The opinion goes on to set out at length the limitations of keyword searching, and the need for sampling of the results of such searches, finding that there was no evidence that the defendant did anything but turn over all documents to the plaintiff that were identified by the keywords used as non-privileged. Later in the opinion, in several lengthy footnotes, Judge Grimm first goes on to describe what alternatives exist to keyword searching (including fuzzy search models, Bayesian classifiers, clustering, and concept and categorization tools), citing the Sedona Search Commentary,[30] and second, provides a mini-law review essay on the subject of whether Judge Facciola's recent opinions in O'Keefe and Equity Analytics should be read to require that expert testimony under Federal Rule of Evidence 702 be presented to the finder of fact in every case involving the use of search methodologies. In Judge Grimm's view:

> "Viewed in its proper context, all that O'Keefe and Equity Analytics required was that the parties be prepared to back up their positions with respect to a dispute involving the appropriateness of ESI search and information retrieval methodology—obviously an area of science or technology—with reliable information from someone with the qualifications to provide helpful opinions, not conclusory argument by counsel. . . . The message to be taken from O'Keefe and Equity Analytics, and this opinion is that when parties decide to use a particular ESI search and retrieval methodology, they need to be aware of literature describing the strengths and weaknesses of various methodologies, such as [the Sedona Search Commentary] and select the one that they believe is most appropriate for its intended task. Should their selection be challenged by their adversary, and the court be called upon to make a ruling, then they should expect to support their position with affidavits or other equivalent information from persons with the requisite qualifications and experience, based on sufficient facts or data and using reliable principles or methodology."[31]

Post-Victor Stanley, a number of other opinions have discussed various aspects of keyword searching and its limitations. For example, one party's attempt to propound 1,000 keywords, and another party's refusal to supply any keywords altogether, led U.S. Magistrate Judge Andrew Peck to lambast the parties for having the case be "the latest example of lawyers designing keyword searches in the dark, by the seat of the pants," and to go on to hold that

> "Electronic discovery requires cooperation between opposing counsel and transparency in all aspects of preservation and production of ESI. Moreover, where counsel are using keyword searches for retrieval of ESI, they at a minimum must carefully craft the appropriate keywords, with input from the ESI's custodians as to the words and abbreviations they use, and the proposed methodology must be quality control tested to assure accuracy in retrieval and elimination of 'false positives.' It is time that the Bar—even those lawyers who did not come of age in the computer era—understand this."[32]

[29] Id. at 256-57.

[30] Id. at 259 n.9.

[31] Id. at 260 n.10.

[32] William A. Gross Construction Assocs., Inc. v. Am. Mftrs. Mutual Ins. Co., 256 F.R.D. 134, 135 (S.D.N.Y. 2009).

The new case law on search and IR amounts to a change in the way things were before, for both the bar and the bench: counsel has a duty to fairly articulate how they have gone about the task of finding relevant digital evidence, rather than assuming that there is only one way to go about doing so with respect to ESI (for example, using keywords), even if the task appears to be a trivial or uninteresting one to perform. Arguably, the "reasonableness" of one's actions in this area will be judged in large part on how well counsel, on behalf of his or her client, has documented and explained the search process and the methods employed. In an increasing number of cases, courts can be expected not to shirk from applying some degree of searching scrutiny to counsel's actions with respect to IR. This may be greeted as an unwelcome development by some, but it comes as an inevitable consequence of the heightened scrutiny being applied to all aspects of e-discovery in the wake of the newly revised Federal Rules of Civil Procedure.

Given decisions such as in Disability Rights, O'Keefe, and Creative Pipe, it seems certain that in a few years' time there will be large and increasing jurisprudence discussing the efficacy of various search methodologies as employed in litigation.[33] Nevertheless, the legal field is still very much a vast tabula rasa awaiting common law development on what constitutes alternative forms of "reasonable" searches when one or more parties are faced with finding "any and all" responsive documents in increasingly vast data sets.

3.2 Keywords, Concepts, and IR Researchers

An IR researcher reading the above discussion may be justifiably concerned at the casual use of technical terminology, and the technical import imputed to casual terminology, in legal rulings on IR. The term "keyword searching," for instance, has been used in the IR literature to refer to any or all of exact string matching, substring matching, Boolean search, or statistical ranked retrieval, applied to any or all of free text terms (e.g., space-delimited tokens or character n-grams), manually assigned uncontrolled terms, or manually or automatically assigned controlled vocabulary terms, with or without augmentation by any combination of stemming, wildcards, multi-word phrase formation, proximity and/or word order restrictions, field restrictions, and/or a variety of other operators.

Thus "keyword searching," while having some implication of more or less direct matching of terms in a query and document, is at best an extraordinarily vague term. Indeed, in the IR research literature it is used almost exclusively in a derogatory fashion, to refer to any method which an author believes is inferior to their preferred technique. No one claims to have built a keyword search system, yet somehow the world is full of them. At the bottom, all IR is based on terms that are often referred to as "keywords" (i.e., words on which an index (a "key") has been built to facilitate retrieval), and the thought that legal penalties might be imposed based on whether someone has or has not used "keyword searching" is therefore alarming.

In contrast, "concept searching" is almost uniformly used with a positive connotation, in both technical and marketing literature. But the breadth of technologies that "concept searching" has referred to includes controlled vocabulary indexing (manual

[33] We note that at least one important decision has been rendered by a court in the United Kingdom, which in sophisticated fashion similarly has analyzed keyword choices by parties at some length. See Digicel (St. Lucia) Ltd. & Ors. v. Cable & Wireless & Ors., [2008] EWHC 2522 (Ch.).

or automatic, with or without thesauri), multi-word phrase formation (by statistical and/or linguistic means), statistical query expansion methods, knowledge representation languages and inference systems from artificial intelligence, unsupervised learning approaches (including term clustering, document clustering, and factor analytic methods such as latent semantic indexing), as well as simple stemming, wildcards, spelling correction and string similarity measures. These technologies have wildly varying, and often poorly understood, behavior. They also overlap substantially the list of technologies that have been referred to (by others) as "keyword searching."

The need for more precise understanding of the usefulness of specific IR technologies in e-discovery is clear, as is the need for much more attention to the overall process in which they are used. We address these issues in the next two sections and then look ahead to next steps in Section 6.

4 Information Retrieval Evaluation

Unlike typical *data retrieval* tasks in which the content of a correct response to a query is easily specified, IR tasks treat the correctness of a response as a matter of opinion. A correctly returned *document* (broadly conceived as any container of information) is considered relevant if the user would wish to see it, and not relevant otherwise. The concept of *relevance* is fundamental, and that therefore is where we begin our review of IR evaluation. An important consequence of relevance being an opinion (rather than an objectively determinable fact) is that retrieval *effectiveness* is a principal focus for evaluation. That is not to say that efficiency is not important as well; just that there would be little value in efficient implementation of ineffective techniques.

The history of IR system development has been shaped by an evaluation-guided research paradigm known broadly as the "Cranfield tradition," which we describe in Section 4.2. As with any mathematical model of reality, evaluation in the Cranfield tradition yields useful insights by abstracting away many details to focus on system design, which is of course just one part of the rather complex process of information seeking that people actually engage in. We therefore conclude this section with a review of approaches to evaluation that involve interaction with actual users of those systems.

4.1 Defining Relevance

The term "relevance" has been used in many different ways by scholars and practitioners interested in helping people to find the information that they need. For our purposes, three of those ways are of particular interest. In *information seeking behavior* studies, "relevance" is used broadly to essentially mean "utility" (i.e., whether a document would be useful to the requestor). In most IR research, and particularly in the Cranfield tradition described below, relevance is more narrowly defined as a relation between a topic and a document. In e-discovery, relevance (or, the more commonly used term in a legal context, "relevancy") is often used as a synonym for "responsiveness," with the literal interpretation "what was asked for." In this section, we begin with a very broad conception of relevance; we then use that background to situate the narrower context used in much of recent IR research.

In a metastudy based on nineteen information seeking behavior studies, Bales and Wang identified fourteen *relevance criteria* that researchers had examined in one or more of those studies (Bales and Wang, 2006). Among the criteria that might be of interest in e-discovery applications were topicality, novelty, quality, recency, stature of the author, cost of obtaining access, intelligibility, and serendipitous utility for some purpose other than that originally intended. Their model of how relevance is decided by users of a system is fairly straightforward: the users observe some attributes of a document (e.g., author, title, and date), from those attributes they form opinions about some criteria (e.g., novelty), and from those criteria they decide whether the document is useful.

This closely parallels the situation in e-discovery, although naturally additional attributes will be important (in particular, custodian), and (in contrast to information seeking by end users) the criteria for "relevancy" are always stated explicitly. That's not to say that those explicitly stated criteria are specified completely, of course—considerable room for interpretation often remains for the simple reason that it is not practical to consider absolutely everything that might be found in any document when specifying a production request.

The narrower conception of relevance in the IR research community arises not from a difference in intent, but rather from a difference in focus. IR research is fundamentally concerned with the design of the systems that people will use to perform information seeking, so evaluation of those systems naturally focuses on the parts of the overall information seeking task that the system is designed to help with. Although factors such as novelty and quality have been the focus of some research (e.g., in the First Story Detection task of the Topic Detection and Tracking evaluation (Wayne, 1998), and in the widely used PageRank "authority" measure, respectively (Brin and Page, 1998)), the vast majority of IR evaluation is concerned principally with just one aspect of relevance: topicality.

Although other definitions have been used, the most widely used definition of *topical relevance* by IR researchers is substantive treatment of the desired topic by any part of the document. By this standard, an email that mentions a lunch with a client in passing without mentioning what was discussed at that meeting would be relevant to a production request asking for information about all contacts with that client, but it would not be relevant to a production request asking for information about all discussions of future prices for some product. Relevance of this type is typically judged on a document by document basis, so the relevance of this one document (in the sense meant by IR researchers) would not be influenced by the presence in the collection of another document that described what was discussed at that lunch.

It is important to recognize that the notion of relevance that is operative in e-discovery is, naturally, somewhat more focused than what has been studied in information seeking behavior studies generally (which range from very broad exploratory searches to very narrowly crafted instances of known-item retrieval), but that it is at the same time somewhat broader than has been the traditional focus of Cranfield-style IR research. We therefore turn next to describe how the "Cranfield tradition" of IR evaluation arose, and how that tradition continues to shape IR research.

4.2 The Cranfield Tradition

The subtleties of relevance and complexities of information seeking behavior notwithstanding, librarians for millennia have made practical choices about how to organize and access documents. With the advent of computing technology in the late 1950's, the range of such choices exploded, along with controversy over the best approaches for representing documents, expressing the user's information needs, and matching the two. Ideally each approach would be tested in operational contexts, on users with real information needs. In practice, the costs of such experiments would be prohibitive, even if large numbers of sufficiently patient real world users could be found.

A compromise pioneered in a set of experiments at the Cranfield Institute of Technology in the 1960's (Cleverdon, 1967) was to represent the basics of an information access setting by a test collection with three components:

- Users are represented by a set of text descriptions of information needs, variously called *topics*, *requests*, or *queries*. (The last of these should be avoided, however, as "query" is routinely used to refer to a derived expression intended for input to particular search software.)
- The resources searched are limited to a static collection of documents.
- Relevance is captured in a set of manually-produced assessments (*relevance judgments*), which specify (on a binary or graded scale) the (topical) relevance of each document to the topic.

The effectiveness of a retrieval approach is then measured by its ability to retrieve, for each topic, those documents which have positive assessments for that topic. Assuming binary (i.e., relevant vs. non-relevant) assessments, two measures of effectiveness are very commonly reported. *Recall* is the proportion of the extant relevant documents that were retrieved by the system, while *precision* is the proportion of retrieved documents which were in fact relevant. Together they reflect a user-centered view of the fundamental tradeoff between false positives and false negatives. These measures are defined for an unordered set of retrieved documents, such as would be produced by a Boolean query, but they were soon extended to evaluate systems which produce rankings of documents (e.g., by defining one or more cutoff points in the ranked list).[34]

The test collection approach greatly reduced the cost of retrieval experiments. While the effort to build a test collection was substantial, once built it could be the basis for experiments by any number of research groups. The relative ease of test collection experiments undoubtedly contributed to the statistical and empirical bent of modern IR research, even during the decades of the 1970's and 1980's when most other work on processing natural language was focused on knowledge engineering approaches.

This test collection model of IR research evolved with relatively minor changes from the 1960's through the 1980's. As new test collections were (infrequently) produced, they grew modestly in size (from hundreds to a few thousands of documents). Even this small increase, however, ruled out exhaustively assessing each document for relevance to each topic. Pooling (i.e., taking the union of the top ranked documents from a variety of ranked retrieval searches based on each topic) and having the assessor judge this pool of documents was suggested as a solution; traditional effectiveness measures could then be computed as if the pooling process had found all of the relevant documents (Spärck Jones and van Rijsbergen, 1975). The strategy was implemented in a minimal fashion

[34] Strictly speaking, unlike for a set, it is not meaningful to refer to the "recall" or "precision" of a *ranking* of documents. The popular ranked-based measures of Recall@k and Precision@k (which measure recall and precision of the set of top-ranked k documents) nominally suggest a recall or precision orientation for ranking, but actually compare ranked retrieval systems identically on individual topics. One can observe the recall-precision tradeoff in a ranking, however, by varying the cutoff k; e.g., increasing k will tend to increase recall at the expense of precision.

for producing several test collections of this era. For instance, the widely used CACM collection used relevance judgments based on the top few documents from each of seven searches, all apparently executed with the same software (Fox, 1983).

A handful of small test collections thus supported the development of the major approaches to statistical ranked retrieval of text, approaches which are now ubiquitous within search engines and a wide range of text analysis applications. By the end of the 1980's, however, a variety of weaknesses in existing test collections were apparent. The size of document collections searched by commercial vendors, government agencies, and libraries had vastly exceeded the size of test collections used in IR research. Further, the full text of documents was increasingly available, not just the bibliographic abstracts in most test collections of the time. Managers of operational IR systems were largely ignoring the results of IR research, claiming with some justification that the proposed methods had not been realistically tested. IR research was itself not in a healthy state, with plausible techniques failing to show improvements in effectiveness, and worries that the field as a whole had overfit to the available test collections. A further nagging problem was variation among researchers in their choices of which topics and documents to use in any given experiment, and in how effectiveness figures were computed, leading to difficulties in comparing even results generated from the same test collection.

The TREC evaluations by the US National Institute of Standards and Technology (NIST) were a response to these problems (Voorhees and Harman, 2005). The test collection produced in TREC's first year (1992) had 100 times as many documents as the typical test collection of the time, with full text instead of bibliographic records. The first TREC evaluation introduced numerous other innovations, including pooling from multiple research groups, a synchronized timetable of data release and result submissions, more detailed descriptions of information needs, relevance assessment by paid professionals using carefully specified procedures, computation of effectiveness measures by a single impartial party using publicly available software, and a conference with attendance restricted to groups participating in the evaluation. Subsequent TRECs have greatly expanded the range of information access tasks studied and the number and size of the data sets used (Voorhees and Harman, 2005; Harman, 2005), and several IR evaluation forums inspired by TREC have emerged around the world (Peters and Braschler, 2001; Kando et al, 2008; Kazai et al, 2004; Majumder et al, 2008).

The large size of the TREC collections led to initial doubts that pooling methods could produce reliable relevance judgments. Studies conducted on the first TREC test collections, with sizes on the order of a few hundred thousand documents, were reassuring because the relative ordering of systems by effectiveness was found to be largely unchanged if different assessors were used or if a particular system's submitted documents were omitted from the pool (Buckley and Voorhees, 2005). The latter result is particularly encouraging, in that it suggested that pools could be used to provide reliable relevance assessments for systems which had not themselves participated in the corresponding TREC evaluation.

Collections continued to grow in size, however, and a study on the AQUAINT collection (over 1 million documents) was the first to find a clearly identifiable pool bias against a particular class of retrieval systems (Buckley et al, 2006). By then, test collections with as many as 25 million documents were in routine use (Clarke et al, 2005), so worries increased. Anecdotal reports also suggested that Web search companies were successfully tuning and evaluating their systems, by then handling billions of documents, using approaches very different from traditional pooling.

These concerns sparked a blossoming of new evaluation approaches for ranked retrieval. Various combinations of the following approaches have been recently proposed:

- After-the-fact unbiasing of biased document pools (Büttcher et al, 2007)
- Using effectiveness measures and/or statistical significance tests that are more robust to imperfect relevance judgments (Buckley and Voorhees, 2004; Sanderson and Zobel, 2005; Yilmaz and Aslam, 2006; Sakai and Kando, 2008; Moffat and Zobel, 2008)
- Treating effectiveness values computed from a limited set of relevance judgments as estimators (with various statistical properties) of effectiveness on the collection (Aslam et al, 2006; Baron et al, 2007; Yilmaz and Aslam, 2006; Carterette et al, 2008)
- Using sampling or pooling methods that concentrate assessments on documents which are most likely to be relevant, representative, revealing of differences among systems, and/or able to produce unbiased estimates of effectiveness (Lewis, 1996; Zobel, 1998; Cormack et al, 1998; Aslam et al, 2006; Baron et al, 2007; Soboroff, 2007; Carterette et al, 2008)
- Leveraging manual effort to find higher quality documents (Cormack et al, 1998; Sanderson and Joho, 2004)
- Using larger numbers of queries, with fewer documents assessed per query (Sanderson and Zobel, 2005; Carterette et al, 2008)

In Section 5 we look at the particular combinations of these techniques that have been brought to bear in the TREC Legal Track.

4.3 Interactive Evaluation

A test collection abstracts IR tasks in a way that makes them affordably repeatable, but such an approach leads directly to two fundamental limitations. First, the process by which the query is created from the specification of the information need (the topic) will necessarily be formulaic if strict repeatability is to be achieved. Second, and even more important, the exploratory behavior that real searchers engage in—learning through experience to formulate effective queries, and learning more about what they really are looking for—is not modeled. In many cases, this process of iterative query refinement yields far larger effects than would any conceivable improvements in automated use of the original topic (Turpin and Scholer, 2006). Accordingly, IR research has long encompassed, in addition to Cranfield-style experiments, other evaluation designs that give scope for human interaction in carrying out retrieval tasks (Ingwersen, 1992; Dumais and Belkin, 2005). In this section, we review some key aspects of interactive evaluations.

Design considerations. In designing an interactive evaluation, it is important to recognize that there is more than one way in which an end user can interact with a retrieval system. In some cases, for example, the end user will interact directly with the system, specifying the query, reviewing results, modifying the query, and so on. In other cases, the end user's interaction with the system will be more indirect; the end user defines the information need and is the ultimate arbiter of whether that need has been met, but does not directly use the retrieval software itself. If an interactive evaluation is to model accurately a real-world task, it must model the mode of interaction that characterizes real-world conditions and practice for that task.

Gauging effectiveness. While incorporating end-user interaction in an evaluation of IR systems can make for a more realistic exercise, it can also make for a more complex (and more resource-intensive) task. There are three reasons for this. First, by introducing the end user into the task, one introduces additional dimensions of variability (e.g., background knowledge or search experience) which can be difficult to control for. Second, by introducing some specific end user as the arbiter of success, standardizing a definition of relevance becomes more challenging. Third, the presence of a user introduces a broader range of measures by which the success of a given retrieval process can be gauged. Apart, for example, from quantitative measures of retrieval effectiveness, such as recall and precision, one may also be interested in measures of learnability, task completion time, fatigue, error rate, or satisfaction (all of which are factors on which the system's likelihood of real-world adoption could crucially depend).

These considerations have resulted in considerable diversity of interactive evaluation designs, each of which strikes a different balance among competing desiderata (Ingwersen and Järvelin, 2005). Interactive evaluation trades away some degree of generalizability in order to gain greater insight into the behavior and experiences of situated users who can engage in a more complex information seeking process than is typically modeled in the Cranfield tradition. It is therefore useful to think of interactive and Cranfield-style evaluation as forming a natural cycle: through interactive evaluation, we learn something about what our systems must do well, through Cranfield-style evaluation we then learn how to do that well, which in turn prompts us to explore the problem space further through interactive evaluation, and so on. In the next section, we describe how these two evaluation paradigms have informed the design of the TREC Legal Track.

5 The TREC Legal Track

In 2006, three of the authors organized the first evaluation in the TREC framework of text retrieval effectiveness for e-discovery: the 2006 TREC Legal Track (Baron et al, 2007). Subsequent TREC Legal Tracks have been organized by some of us (and others) in 2007 (Tomlinson et al, 2008), 2008 (Oard et al, 2009) and 2009 (Hedin et al, 2010), with another planned for 2010.

As with all TREC tracks, we have sought to attract researchers to participate in the track (in our case, both from industry and academia), develop guidelines both for participating research teams and for relevance assessors, manage the distribution of large test collections, gather and analyze results from multiple participants, and deal with myriad technical and data glitches. Other challenges have been unique to the Legal Track. The biggest ones flow from the fact that we have two audiences for our results: IR researchers whose efforts we hoped to attract to work on e-discovery problems, and the much larger legal community for whom we hoped the TREC results would provide some measure of insight and guidance. Making our results compelling to the legal community required that queries be provided with substantial context in the form of simulated legal complaints (as discussed in Section 5.3 below). Documents similar to those encountered in e-discovery were desirable as well; Sections 5.1.1 and 5.1.2 introduce the test collections we used.

Attempting to capture all the important aspects of an e-discovery setting in a single simulation would likely lead to a task too expensive to run and too complex to attract researcher interest. We instead designed three tasks that measure the effectiveness of different aspects of search processes and IR technology (Sections 5.2 to 5.4).

The particular nature of relevance in the e-discovery context led us to believe that relevance assessments should be carried out by personnel with some legal background. This has required recruiting and training literally hundreds of law students, paralegals, and lawyers as volunteer relevance assessors, as well as the creation of two Web-based systems to support a distributed assessment process.

A final challenge is that good retrieval effectiveness in an e-discovery context means high recall (i.e., that the vast majority of relevant documents must be retrieved). Most IR evaluations, and in particular recent work with an eye toward Web search engines, have focused most strongly on precision (or more specifically, on the presence of relevant documents) near the top of a ranked list. This focus has affected not just the choice of effectiveness measures, but also how topics were selected and how documents were chosen for assessment. In particular, ad hoc evaluations at TREC have typically assumed that pooling the top-ranked 100 documents from each participating system would cover most of the relevant documents for each topic, which we quickly learned was inadequate for the scale of the test collections and topics in the Legal Track. Hence the evaluation approach had to be rethought for the Legal Track, with the result that stratified sampling and corresponding estimation methods have played a larger role in the Legal Track (see Sections 5.2 and 5.3) than in previous TREC tracks.

5.1 Test Collections

Two test collections have been used in the TREC Legal Tracks. Each captures some aspects of current e-discovery settings, while missing others. Because they include documents of types not previously available in IR test collections, the collections are also likely to be of interest to IR researchers working on those specific document types.

5.1.1 The IIT CDIP Collection

Our first test collection, used for all tasks except the 2009 Interactive task, was the *Illinois Institute of Technology Complex Document Information Processing Test Collection, version 1.0*, referred to here as “IIT CDIP” and informally in the TREC community as the “tobacco collection.” IIT CDIP was created at the Illinois Institute of Technology (Lewis et al, 2006; Baron et al, 2007) and is based on documents released under the Master Settlement Agreement (MSA) between the Attorneys General of several US states and seven US tobacco companies and institutes.[35] The University of California San Francisco (UCSF) Library, with support from the American Legacy Foundation, has created a permanent repository, the Legacy Tobacco Documents Library (LTDL), for tobacco documents (Schmidt et al, 2002), of which IIT CDIP is a cleaned up snapshot generated in 2005 and 2006.

IIT CDIP consists of 6,910,192 document records in the form of XML elements. Records include a manually entered document title, text produced by Optical Character Recognition (OCR) from the original document images, and a wide range of manually created metadata sub-elements that are present in some or all of the records (e.g.,

[35] http://ag.ca.gov/tobacco/msa.php

sender, recipients, important names mentioned in the document, controlled vocabulary categories, and geographical or organizational context identifiers).

IIT CDIP has strengths and weaknesses as a collection for the Legal Track. The wide range of document lengths (from 1 page to several thousand pages) and genres (including letters, memos, budgets, reports, agendas, plans, transcripts, scientific articles, email, and many others), as well as the sheer number of documents, are typical of legal discovery settings. The fact that documents were scanned and OCRed, however, is more typical of traditional discovery than e-discovery settings. The fact that the MSA documents were the *output* of discovery proceedings raised questions as to their appropriateness as *input* to TREC's simulation of a legal discovery situation. This concern was mitigated to some extent by the fact that the MSA documents resulted from hundreds of distinct document requests in multiple legal cases. We further addressed this concern by using a range of topics in the Legal Track experiments, some with content highly similar to MSA discovery requests, and others very different.

One disappointment was that the OCR errors in the IIT CDIP collection occupied so much of participants' attention, and indeed appeared to discourage participation. An early TREC track investigated retrieval from simulated OCR text (Voorhees and Garofolo, 2005), and while effectiveness was shown to be reduced by OCR noise, it did not leave the impression that OCR errors were a huge issue. At the scale of the IIT CDIP collection, however, OCR errors inflate vocabulary size to such an extent that the use of standard word-based indexing methods was no longer practical for some participants. Thus Legal Track participants had to spend time dealing with systems issues that, while interesting, were not particularly relevant to e-discovery.

5.1.2 The TREC Enron Collection

Email is without question the *sine qua non* of present practice in e-discovery. Development of an email test collection was therefore important if we were to support representative evaluation for existing e-discovery tools, and perhaps to foster the development of a new generation of tools. The IIT CDIP collection actually contains a substantial number of email messages (estimated at about 10% of the collection), but the printing, scanning, and OCR process makes recovery of the content of those emails, and particularly the important metadata in their headers, difficult. A second alternative would have been to use a collection of email messages from the Enron Corporation that had been obtained by MIT in 2004 from Aspen Systems, a company acting on behalf of the Federal Energy Regulatory Commission (FERC). Researchers at MIT, SRI International, and Carnegie Mellon University (CMU) worked together to make this collection of about 250,000 email messages (after de-duplication) freely available on the Internet from CMU (for this reason, we refer to this collection as the "CMU Enron collection"). This resulted in an explosion of research using the collection in a large number of applications, including internal use of the collection by a substantial number of e-discovery vendors. The CMU Enron collection has one key limitation, however: it contains no attachments. This is problematic for e-discovery because present practice requires that decisions about responsiveness be reached at the level of a "record," which in the case of email is typically interpreted as a message together with its attachments.

We therefore elected to create a new Enron email collection that includes attachments for use in the TREC Legal Track. There is no single "Enron collection" for the simple reason that Aspen Systems distributes all available messages (with their attachments) on whatever date the collection is requested. On some occasions, messages have been withdrawn from the collection by FERC for specific reasons (e.g., ongoing litigation). On other occasions, messages have been added to the collection (e.g., as additional materials became available from FERC). The net effect is that more emails are now available than were available at the time that the CMU Enron collection was created.

Our new collection, which we call the "TREC Enron collection" contains 569,034 messages after de-duplication. We employed an existing commercial processing pipeline[36] to convert the format that had been received from Aspen Systems into a format similar to the mbox standard used by Unix mail handling systems (which closely resembles the format of the messages in the CMU Enron collection). Metadata extracted from the headers was also provided using an XML format developed by the Electronic Discovery Reference Model project (see Section 6.2) to support interchange of processed collections among heterogeneous systems.

The attachments received from Aspen Systems were in their original format (e.g., Microsoft Word, PDF, or JPEG); where possible, ASCII text was extracted from each attachment using the same commercial processing pipeline as had been used for processing the email messages. The distributed version thus contains each message in the "native" form received from Aspen Systems, and also (where possible) in a more easily handled text-only form. A third form, document images, was also created using the same commercial processing pipeline. This avoided the need for the relevance assessment platform to deal with multiple content formats: it simply displayed an image of each page to the assessor. The document image files are fairly large, so they are not normally distributed with the collection. We consider the collection used in 2009 to be a beta release. A number of issues (most notably de-duplication) are still being worked through, and it is expected that an improved collection will be used for the 2010 Legal Track.

5.2 The Interactive Task

There are, in the course of a lawsuit, a wide range of tasks to which an attorney can usefully apply IR methods and technologies. Early in the lawsuit, an attorney may conduct exploratory searches of a document collection in order to test hypotheses and to build a theory of the case; further along, an attorney may search for documents relevant to the activities of a particular individual in order to prepare for the deposition of a witness; as trial approaches, an attorney may search through the set of generally relevant documents in order to find the small subset that he or she would like to enter as exhibits in trial; and so on. The particular task that the Legal Track's Interactive task (and, indeed, all tasks in the TREC Legal Track to date) models is *responsive review*: the circumstance in which a party to litigation is served, by the opposing party, with a request for production of documents and, in order to comply with the request, must find and produce, to an extent commensurate with a reasonable good-faith effort, any and all documents that are responsive to the request. In this section, we review key elements of the design of the Interactive Task and summarize results from the first

[36] Clearwell Systems obtained the collection from Aspen Systems and performed the processing described in this section.

two years in which the task was run, 2008 (Oard et al, 2009) and 2009 (Hedin et al, 2010).[37]

Modeling the conditions and objectives of a responsive review, the Interactive task provides a mock complaint and an associated set of requests for production; for purposes of the exercise, each of the requests serves as a separate topic. The task also features a role, that of the "Topic Authority," (TA) modeled on the part played by the senior attorney who would be charged with overseeing the responding party's reply to the request for production and who, in keeping with that role, would form a conception of what sort of material was and was not to be considered responsive to the request. The role of a team participating in the task is modeled on that of a provider of document review services who has been hired by the senior attorney to find the documents that match his or her conception of responsiveness; as such, the goal of a participating team is to retrieve from the test collection all (and only) the documents that are consistent with the TA's definition of what is responsive to the given request. The task is designed to measure how closely each team comes to meeting that goal by estimating the recall, precision, and F_1 (the balanced harmonic mean of recall and precision) attained for a topic by a team.

The senior attorney must eventually certify to the court that their client's response to the request is (commensurate with a reasonable good-faith effort) accurate and complete. In keeping with that role, the TA takes into account both considerations of subject-matter relevance and considerations of legal strategy and tactics (e.g., to what extent a broad interpretation of responsiveness, one that reduces the likelihood of a challenge for underproduction, would serve the client better than a narrow interpretation of responsiveness, one that reduces the likelihood of producing potentially damaging material that arguably could have been held back) in arriving at a conception of what should and should not be considered responsive to the request (topic). Each topic has a single TA, and each TA has responsibility for a single topic.

In the Interactive task, the TA's role is operationalized in three ways. The first is topic clarification; the TA acts as a resource to which teams can turn in order to clarify the scope and intent of a topic. A team can ask for up to 10 hours of a TA's time for this purpose. While we instruct the TAs to be free in sharing the information they have about their topics, we also ask that they avoid volunteering to one team specific information that was developed only in the course of interaction with another team. The second is review oversight. In order to be able to obtain valid measures of effectiveness, it is essential that the samples be assessed in accordance with the TA's conception of relevance; it is the role of the TA to provide assessors with guidance as to what is and is not relevant. The third is final adjudication. As discussed below, teams are given the opportunity to appeal specific relevance assessments they believe have been made in error (i.e., inconsistently with the TA's conception of responsiveness); it is the role of the TA to render a final judgment on all such appeals.

When an attorney vouches for the validity of a document production, he or she is vouching for the accuracy of a binary classification of the document population implicated by the request: a classification of the population into the subset that is responsive to the request for production and the subset that is not. When an e-discovery firm supports an attorney in this effort, it must make a similar responsiveness determination. The Interactive task, modeling this requirement, specifies that each participant's final

[37] A pilot Interactive task was run in 2007 (Tomlinson et al, 2008), but with a very different task design.

deliverable be a binary classification of the full population for relevance to each target topic. Teams are of course free to use relevance ranking as a means of arriving at their result sets, but the final deliverable is a single binary classification (relevant/not relevant) of the full population of documents.

Effectiveness measures are computed using a stratified sampling design that enables us to obtain unbiased estimates for precision and recall. For each topic, the population is stratified on the basis of team submissions. One stratum contains the documents all teams submitted as responsive, another contains documents no team submitted as responsive, and other strata are subsets representing various conflicting submissions. In sampling from those strata, we oversample strata that are likely to contain more responsive documents, as they have the most impact on system comparisons. This inevitably means that very large strata, and in particular the "All-Negative" stratum, are under-represented in the sample, and so our estimates of responsiveness in those strata have higher variance. This is, however, both inevitable to a degree given the low frequency of responsive documents in these strata, and at least an improvement over the ignoring of these documents in traditional pooling.

We use a two-stage procedure to assess the sample for relevance. In the first stage, volunteer assessors, under the guidance of the TA, conduct a complete first-pass review of the sampled documents. Assessors are provided with topic-specific guidelines that capture all guidance that the TA had provided to any team and with access to the TA for any further clarification that they feel is needed. In the second stage, teams are given access to the results of the first-pass assessment of the sample and invited to appeal any assessments that they believe are inconsistent with specific guidance that they received from the TA. The TA personally renders a final judgment on all appealed assessments. These final, post-adjudication, assessments are the basis for estimating recall, precision, and F_1; see the appendix to 2008 Legal Track overview paper for details (Oard et al, 2009).

For the 2008 Interactive task, the test collection was the IIT CDIP collection of scanned business records (see Section 5.1.1). The complaint was a realistic but fictitious securities class action; associated with the complaint were three production requests (topics). By way of illustration, one topic (Topic 103) called for:

> All documents which describe, refer to, report on, or mention any "in-store," "on-counter," "point of sale," or other retail marketing campaigns for cigarettes.

As with real-world document requests, the topic, as stated, leaves scope for interpretation (how, for example, do you define a "marketing campaign" or, more specifically, a "retail marketing campaign") and that is where the TA's approach to the request is decisive.

Four teams (two academic and two commercial) participated in the 2008 Interactive task. Also, a pool formed from the 68 runs submitted to the 2008 Ad Hoc task (see Section 5.3) which provided ranked retrieval results to a maximum depth of 100,000 (generated without any TA interaction for all three Interactive task topics) was included as an additional result set (the "Ad Hoc Pool").

The results for Topic 103 proved to be the most informative, as that is the one topic for which all teams submitted results, and it was the only topic for which any team made extensive use of the appeal and adjudication process. We observed considerable variation in the amount of time with the TA that the teams used, from just 5 to 485 of the available 600 minutes. There was also considerable variation in the number of documents that teams submitted as relevant, from 25,816 to 608,807 of the 6,910,192 documents in the collection. On the basis of the adjudicated sample assessments, we estimated that there are 786,862 documents (11.4% of the collection) relevant to Topic 103 in the test collection (as the topic was defined by the TA). All four teams attained quite high precision; point estimates ranged from 0.71 to 0.81. One team (notably the one that made the most use of TA time) attained relatively high recall (0.62), while the other three (all making significantly less use of TA time) obtained recall values below 0.20.

From this 2008 exercise, we learned that it is possible to attain a reasonably high level of recall without sacrificing precision, at least for the one topic for which we have results from several teams. We also found that the appeal and adjudication process resulted in a change in the assessment for a significant number of documents. Of the 13,500 first-pass assessments (for all three topics), 966 were appealed to a TA (almost all from Topic 103). Of these, 762 (79%) were decided in favor of the appealing team (overturning the first-pass assessment). The impact of the appeal and adjudication process was generally an across-the-board improvement in recall, precision, and F_1 for all teams; in particular, adjudication did not result in a change in the relative ordering of submitted results by F_1. An additional lesson from 2008 was that running the Interactive task with four teams provided an insufficient number of data points to reach any firm conclusions (on, for example, the possible correlation between the amount of time spent with the TA and the levels of recall and precision achieved).

For the 2009 Interactive task (Hedin et al, 2010), we developed a new test collection from Enron emails (see Section 5.1.2). The complaint was again a realistic but fictitious securities class action. Associated with the complaint were seven production requests (topics); the focus of these requests ranged from off-balance sheet accounting to gambling on football. Topic 202, for example, called for:

> All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in transactions that the Company characterized as compliant with FAS 140 (or its predecessor FAS 125).

Topic 207 called for:

> All documents or communications that describe, discuss, refer to, report on, or relate to fantasy football, gambling on football, and related activities, including but not limited to, football teams, football players, football games, football statistics, and football performance.

Eleven teams (three academic and eight commercial, together representing three countries) participated in the 2009 Interactive Task. There were no Ad Hoc submissions to pool in the 2009 Interactive task.

As was the case in 2008, we saw considerable variation in the amount of their allotted time with the TA that teams used. For Topic 207, for example, one team used 0 minutes, while another used 295 minutes; for Topic 204 (on document deletion and retention), one team used 0 minutes and another used 500 of their allotted 600 minutes. Also as was the case in 2008, we saw considerable variation in the number of documents that teams submitted as relevant. For Topic 207, for example, one team submitted 5,706 messages as relevant, while another submitted 25,404 of the 569,034 messages in the test collection as relevant. The appeal and adjudication process was more widely utilized in 2009 than it was in 2008: of the seven 2009 topics, six had at least 200 first-pass assessments appealed and one had 967 assessments appealed.

The post-adjudication results for the 2009 topics showed some encouraging signs. Of the 24 submitted runs (aggregating across all seven topics), 6 (distributed across

5 topics) attained an F_1 score (point estimate) of 0.7 or greater. In terms of recall, of the 24 submitted runs, 5 (distributed across 4 topics) attained a recall score of 0.7 or greater; of these 5 runs, 4 (distributed across 3 topics) simultaneously attained a precision score of 0.7 or greater. Further discussion of the results for the 2009 Interactive task can be found in the 2009 track overview (Hedin et al, 2010).

5.3 The Ad Hoc Task

The Ad Hoc task was designed to compare single-pass (or first-pass) automatic search approaches to e-discovery. It was run in the first three TREC Legal Tracks (Baron et al, 2007; Tomlinson et al, 2008; Oard et al, 2009). For each year's task, approximately 50 new topics were created. Each topic was based on one of 3 to 5 realistic but fictitious complaints; for example, complaint 2007-B described a patent infringement action on a device designed to ventilate smoke. Each topic further included a specific request for documents to produce; for example, topic 74 requested

> All scientific studies expressly referencing health effects tied to indoor air quality.

In each year, the set of documents to search was the IIT CDIP collection (see Section 5.1.1 for details).

For comparison to the primarily automatic "ranked" approaches (i.e., approaches based on relevance-ranking formulas) that were anticipated from the TREC participants, a reference Boolean query[38] was created for each topic by two lawyers conducting a simulated negotiation. For example, the negotiation for topic 74 was as follows. The Boolean negotiation started with a fairly narrow initial proposal from the defendant, "`"health effect!" w/10 "air quality"`". The plaintiff responded with a much broader query, "`(scien! OR stud! OR research) AND ("air quality" OR health)`". The negotiation concluded with a mutually agreeable final Boolean query, "`(scien! OR stud! OR research) AND ("air quality" w/15 health)`". The negotiations for other topics followed a similar pattern. The Boolean negotiation was meant to represent one conception of possible "best practice" (in contrast to the currently more common present practice of searching using single keywords).

The evaluation approach evolved over the three years as experience was gained. In 2006, the evaluation was conducted much like a traditional ad hoc TREC task. In particular, the systems were required to rank the documents rather than just return a set. Submitted results were limited to 5,000 documents per topic, and the pool of documents to judge was primarily based on the top-ranked 100 documents from a single submission from each of the 6 participating research groups. A run manually created by an expert interactive searcher and a sampling of the final negotiated Boolean query run were also included in the pool. The production requests (topics) had deliberately been made narrow in hopes of limiting the number of relevant documents to approximately a few hundred so that the pools (averaging 800 documents) could provide good coverage of the relevant documents. The results showed that although many relevant documents were missed by the negotiated final Boolean query, no ranked retrieval system had an average effectiveness (precision at R, averaged over all topics) substantially

[38] As is common, we use "Boolean query" somewhat loosely to mean queries built using not just the three basic Boolean operators (and, or, not), but also truncation and (unordered) proximity operators.

higher than that of the average effectiveness of the final Boolean queries. Both the final Boolean queries and the ranked retrieval systems were outperformed by the expert interactive searcher, despite several limitations imposed for reasons of practicality on the interactive search process (most notably, the expert was limited to submitting 100 documents per topic). Some early experiments with sampling also suggested that the pool's coverage of the relevant documents was disconcertingly low (Tomlinson, 2007).

In 2007, a deeper sampling scheme was introduced. Submitted result sets of as many as 25,000 documents per topic were allowed, and 12 participating research teams submitted a total of 68 result sets, all of which were pooled, producing a set of approximately 300,000 documents per topic. The sampling scheme was used to select between 500 and 1000 documents from the pool to judge for each topic. This sampling process suggested that there were typically remarkably large numbers of relevant documents in the pool (with point estimates averaging about 17,000 relevant documents per topic) even though the production requests had again deliberately been rather narrowly crafted. The sampling also suggested that the negotiated final Boolean query matched just 22% of the relevant documents in the pool (i.e., an average recall across all topics no higher than 0.22), and its precision averaged just 0.29. However, the final Boolean query was still estimated to have a higher precision and recall (when averaged over all topics) than any submitted ranked retrieval run when evaluated at a depth equal to the number of documents returned by the final Boolean query. Post hoc analysis found, however, that when the ranked retrieval systems were measured at depth R for each topic (where R was the estimated number of relevant documents for that topic from the human relevance assessments), several systems achieved a higher F_1 score than the negotiated final Boolean query run. This observation motivated the introduction of a set-based evaluation measure the next year.

In 2008, set-based evaluation became a principal focus of the ad hoc task with the introduction of a requirement that each system specify an optimal retrieval depth K for each topic, representing the depth at which it believed the F_1 measure would be maximized. The focus on relatively narrowly scoped topics was eliminated, and the allowed submission set size was further increased to 100,000 documents per topic. With this more representative range of topics, the relative order between the negotiated final Boolean query run and the submissions from participating teams reversed, with 7 of the 10 participating teams each submitting at least one run with a higher mean recall than the negotiated final Boolean query run (when compared at the number of documents returned by that negotiated final Boolean query run). Comparison with 2007 was facilitated by the fact that the highest scoring submission in 2008 used the same automated approach as a submitted run from 2007. The topics in 2008 were indeed broader, averaging five times as many relevant documents as in 2007, although the Boolean queries matched an average of eight times more documents in 2008 than 2007. Of course, Boolean queries are the product of a human activity, and it is certainly possible that the differences that we saw between 2007 and 2008 are within the natural range of variation. When evaluated with the new set-based evaluation measure, only 5 of the 10 teams submitted a run of higher mean F_1 at depth K than the negotiated Boolean query run, suggesting that picking the threshold K was challenging for the ranked retrieval systems. The top average F_1 at depth K score was just 0.2 (and rarely above 0.4 for individual topics) indicating that either precision or recall (or both) was relatively low for every system.

5.4 The Relevance Feedback Task

The Relevance Feedback task was designed to evaluate the second pass of multi-stage search approaches for e-discovery. The task used a selection of topics from previous years' Ad Hoc or Interactive tasks. Systems were provided with all available relevance judgments from use of the same topic in a previous year, typically consisting of 500 or more documents that had been marked relevant or not relevant. Evaluation of the relevance feedback systems was based entirely on a new set of relevance assessments, usually from a different assessor. This task was run for three years starting in the second year of the TREC Legal Track (Tomlinson et al, 2008; Oard et al, 2009; Hedin et al, 2010).

In 2007, 3 teams submitted a total of 8 runs (only 5 of which actually used the previous relevance judgments in some way) and 10 topics were assessed. In 2008, 5 teams submitted a total of 29 runs (19 of which used the previous relevance judgments) and 12 topics were assessed. Interestingly, the runs that used a prior year's relevance judgments did not generally achieve higher F_1 values on previously unassessed documents (so-called "residual" evaluation) than those that did not. It may be that relevance feedback is particularly challenging with the IIT CDIP collection, which has a fairly high OCR error rate and some fairly long documents. (OCR errors are challenging for some relevance feedback techniques because they create many rare, mostly useless terms, interfering with length normalization and query expansion (Singhal et al, 1995; Taghva et al, 1996).) The fact that a different assessor produced the new relevance judgments is also a potential confounding factor; a small assessor consistency study in 2008 found that documents judged relevant by the original assessor were only judged relevant 58% of the time by a second assessor (Oard et al, 2009).

In 2009, renamed generically as the Batch task (since either first-pass ad hoc or second-pass relevance feedback techniques could be tried), relevance feedback systems were provided with larger sets of relevance judgments (most notably for Topic 103, for which 6,500 adjudicated relevance judgments from the 2008 Interactive task were available, 2,981 of which were assessed as relevant). The submission limit was increased to 1.5 million documents per topic to allow higher values for K to be reported by systems as a basis for set-based measures. In prior years, only documents returned by at least one system had been sampled (thus omitting documents not found by any system from the computation of recall) and typically only 500 relevance assessments were performed for each topic. In 2009, the entire collection was sampled, and the number of documents judged per topic was increased to typically 2,500 in hopes of providing both a more accurate evaluation of the participating runs and an improved potential for comparably evaluating new systems in the future. Four teams submitted a total of 10 runs. On Topic 103 some relevance feedback systems obtained a higher value for F_1 (up to 0.57) than the highest-scoring Interactive task run submitted the previous year (0.51). On another topic (Topic 51), however, all of the automated systems scored less than 0.01 for F_1 despite seemingly ample numbers of training examples (1,361 relevance judgments, 88 of which were assessed as relevant). This troublesome topic highlighted a problem that prevents our present evaluation framework from being reliably used to evaluate recall for topics with a very small proportion of relevant documents. When sampling from 7 million documents, even a small assessor false positive rate (e.g., 1 in 500) on a sample from a sparsely sampled stratum would lead us to estimate the presence of thousands more relevant documents than actually exist, dominating the computation of recall. As the actual number of relevant documents increases, the impact of this problem naturally diminishes.

5.5 The Legacy of the TREC Legal Track

The TREC Legal Track has been a rather humbling experience for IR researchers (including some of the organizers) who expected that modern ranked retrieval techniques would quickly dominate more basic techniques such as manually constructed queries using Boolean and proximity operators. That it took several years to achieve this perhaps says more about what we needed to learn about evaluation design than it does about specific techniques, however. Our thinking about sampling and the design of evaluation measures has evolved considerably over this period, and that will likely be one of the most lasting legacies of the TREC Legal Track. Two advances seem particularly notable in this regard: (1) estimation of recall based on deep sampling, and (2) modeling the full range of topic diversity, including topics that generate far higher yields than are usually studied in IR evaluations. We have also gained important experience with estimating the accuracy of our measurements, although more remains to be done before we will understand the full implications of our sampling designs for future use of our collections by research teams that did not contribute to the set of runs from which the original samples were drawn.

The Legal Track was humbling also for the insight it has provided the IR research community into the immense investment being made in manual review in day-to-day operational e-discovery settings. While the scope of the TREC relevance assessment process was large by TREC standards (in geography, technology, personnel, and data set preparation), the total number of documents assessed across all four years of the Legal Track has been far smaller than the number routinely assessed in even a single real world e-discovery project. Investment at this scale also offers new opportunities. For example, it would be useful to bring an entirely manual review effort into the evaluation (not as assessors, but as a participating team) as a reference point against which to gauge automated efforts.

One of the most lasting legacies of any TREC track is the research community that coalesces around new research problems. The Legal Track has been particularly rich in this regard, drawing IR researchers, e-discovery service providers, law firms and law schools into a continuing dialog about the challenges of IR, and evaluation of IR, in e-discovery applications. The results of that dialog to date (including this paper, with authors from three of those groups) have been narrowly focused on techniques for automating and evaluating responsive review, but it seems reasonable to anticipate that continued interaction will lead to even richer collaboration on a broad range of issues (e.g., establishing evidentiary best practices for responsive review, partially automating privilege review, or rapidly making sense of vast quantities of material received through an e-discovery process). This type of collaboration might also help to reign in some of the more audacious claims now being made in the e-discovery community regarding the efficacy of certain techniques. While we have indeed seen results from the TREC Legal Track that are impressive for their efficacy relative to established baselines, we continue to see marketing claims that seem well beyond what we are presently able to measure, and perhaps even somewhat beyond what is actually possible.

The specific Interactive task design that we used is new to large-scale IR evaluation in two important ways. In earlier work, intermediated search (e.g., search by a librarian on behalf of a patron) was most often studied in the wild (i.e., by observing several episodes of actual information seeking behavior). Scenario-based structured evaluation of interactive search has been, for the most part, focused on settings in which single users work alone (or, in recent work, in which small teams of searchers collaborate). By modeling the role of the senior attorney with a single TA, we gained in both task fidelity (e.g., facilitating system designs based on active learning techniques from machine learning) and assessor consistency.

Progress in research often raises as many questions as it answers, so new issues illuminated by our work and the work of others in the Legal Track are an important legacy as well. Perhaps the most important of these questions is how best to evaluate recall for low-yield topics. Our thinking on evaluation of interactive search has also evolved, from an initial focus on retrieval effectiveness to a more nuanced view that interaction defines a continuum and that the central issue is therefore the balance between cost and effectiveness. In future Interactive evaluation, measuring effort with sufficient fidelity therefore may well be as important as high-fidelity measurement of effectiveness. More work also remains to be done on developing some form of typology of production requests, with an eye towards better understanding of what makes some production requests more challenging than others. There has been considerable work on this question in a more general IR context (Carmel et al, 2006), but the structure present in e-discovery may offer new opportunities to extend that work in interesting directions.

Of course, another legacy of the Legal Track will be the results achieved by the research teams. Batch evaluation has shown that in a substantial number of cases, statistical ranked retrieval techniques can result in more comprehensive search for responsive review than even a relatively sophisticated “best practice” model in which lawyers negotiate queries with Boolean, proximity, and truncation operators without (at the time the queries are negotiated) having access to the collection. The TREC 2008 paper from the University of Waterloo is a good exemplar of this line of research (Lynam and Cormack, 2009). Interactive evaluation has also contributed new insights, most notably that coupling high precision with high recall is possible, at least for some topics. A set of papers published in the 2009 IEEE Conference on Systems, Man and Cybernetics by authors from H5 and elsewhere offer insight into how this was done for one topic in the 2008 Interactive task (Bauer et al, 2009; Hedin and Oard, 2009).

Finally, the Legal Track’s two test collections provide for a legacy that may well extend well beyond e-discovery. Earlier work at TREC had studied the effect of simulated OCR errors on IR, but the IIT CDIP collection is the first large test collection to include naturally occurring OCR. With relevance judgments now available for more than 100 topics, this collection is well suited to support new research on OCR-based IR. The presence of extensive metadata in this collection also offers scope for exploring other important research questions, since scanned documents (e.g., books) are often accompanied by extensive metadata. Although the TREC Enron collection is far less well developed, with only 7 topics at present, we expect that the TREC 2010 Legal Track will extend that collection in ways that will make it of continuing value to researchers who are interested in working with email. Although the CMU Enron collection has been available for some time now, we expect the availability of attachments and of topics with associated relevance judgments will facilitate new lines of research that would otherwise not have been practical.

6 Thinking Broadly

TREC is, of course, just one way of thinking about evaluation for e-discovery. In this section, we briefly address three other perspectives that further enrich the range of available evaluation options.

6.1 Other Study Designs for Responsive Review

Evaluation design is, in the end, always based on a cost-benefit tradeoff. The prodigious amount of manual effort being invested in manual review in ongoing e-discovery proceedings therefore offers the possibility of basing evaluation of automated systems on the results of an actual manual review. This is an appealing option both because the *found data* (i.e., the existing relevance judgments from the manual review process) is essentially free, and because the use of a real case rather than a fictitious complaint removes one possible concern to the fidelity of the evaluation. On the other hand, using real data raises several challenges, including whether the process can be adequately instrumented (e.g., to measure inter-annotator agreement) when it must be performed under significant time pressure and resource constraints, and whether the resulting evaluation resources can be shared with other researchers (which is important both for new work by others, and for replication of reported results).

All of these benefits and challenges are illustrated in a recent study by the Electronic Discovery Institute (Oot et al, 2010; Roitblat et al, 2010) using exhaustive relevance judgments for more than 1.8 million documents that were assessed for responsiveness as part of what is known as a “Second Request” in antitrust law. The documents were originally reviewed for responsiveness in connection with a Department of Justice investigation of the acquisition of MCI by Verizon. In the study, the agreement between two sets of independent assessors and the original review was measured on a sample of 5,000 documents and found to be substantially lower than is typical for TREC evaluations. Two automated systems were then built and run on the full set of 1.8 million documents, yielding results (when scored with the original exhaustive assessments) that were comparable to the results reported for the other assessor’s agreement with the original assessments. The low degree of agreement between the original and subsequent manual reviews makes it somewhat difficult, however, to draw clear conclusions about the effectiveness of automated classification from the published results.

Although we are not yet aware of other studies using this type of found data, we believe that this approach offers an important complement to the insights that we can derive from TREC-style evaluations by bringing in results from actual e-discovery practice. Moreover, over time we expect that it will be both possible and highly desirable to better instrument the original review process, thus further extending the range of insights that can be gained through such an approach. If concerns about future use by others of the resulting test collections can be resolved (e.g., through licensing arrangements that restrict redistribution, or through innovative approaches such as the “algorithm submission” that has been used in TREC to evaluate email spam filters on sensitive content (Cormack and Lynam, 2006)), then this way of developing test collections eventually could and should replace the somewhat less realistic and considerably more expensive process that we have used in the TREC Legal Track to date.

6.2 Looking Beyond Search

Concurrent with the evolution of the TREC Legal Track, a more comprehensive industry working group has coalesced around an initiative known as the "Electronic Discovery Reference Model" (EDRM), whose stated purpose is"to develop guidelines and standards for e-discovery consumers and providers."[39]

EDRM complements the more limited scope of responsive review evaluations such as the TREC Legal Track and the Electronic Discovery Institute study described above by articulating a "lifecycle of e-discovery" consisting of nine broad task categories: information management; identification; preservation; collection; processing; review; analysis; production; and presentation. Virtually all of these broad task categories could benefit from tailored application of IR or closely related technologies (e.g., text classification), some quite directly, and many of them raise interesting and important evaluation questions. For example, records management professionals seek to organize business records in ways that support both ongoing needs of the organization and possible future uses such as e-discovery. The EDRM identification task category includes a close analogue to the collection selection task in distributed IR research (albeit one that in e-discovery depends more on selecting custodians than has typically been the focus in research to date on distributed IR). De-duplication (to reduce effort and improve consistency) is a step in the EDRM processing task category that would benefit substantially from well-designed task-based models of document similarity. Sorting into broad categories (e.g., to avoid devoting review effort to email spam or pornography) is another step in the EDRM processing task category in which text classification techniques are needed. EDRM's analysis task category includes "early case assessment," which could benefit from exploratory search such as clustering and link analysis (Solomon and Baron, 2009). Privilege review (in the EDRM review task category along with review for responsiveness) is another step in which text classification could be useful. And, of course, opportunities abound for productive use of IR techniques in the EDRM presentation task category in which the party receiving the documents then seeks to make sense of them and to make use of them.

In setting up the TREC Legal Track, there was an obvious relationship between TREC's "Cranfield tradition" and the idea that lawyers are engaged in an IR process (Baron, 2007). As EDRM makes clear, while this is most certainly true, the need for IR extends well beyond responsive review, and indeed well beyond Cranfield-style evaluation. Some of these related IR technologies have been evaluated in other contexts in TREC in the past (e.g., in the TREC interactive, filtering, and spam tracks), but evaluation in an e-discovery context remains a future task. The EDRM project is evolving a rich set of guidelines and standard reference collections for many of these tasks, and as that process evolves it seems reasonable to expect that the work we have done in the TREC Legal Track to date will one day be subsumed in some more comprehensive effort.

6.3 Certifying Process Quality

Evaluation-guided research is to some extent a two-edged sword. On the one hand, it fosters convergence on a set of well modeled challenge problems for which insightful

[39] http://edrm.net/

and affordable evaluation resources can be constructed. On the other hand, the mere existence of those resources tends to reinforce that convergence, regardless of whether what was originally modeled remains the best approach. Of course, the marketplace, despite sometimes inflated expectations and exhortations, does act to counterbalance any tendency of the research community to focus its attention too narrowly. Ultimately, what is needed is some way of balancing these two forces in a manner that maximizes synergy and minimizes wasted effort. One broad class of approaches that has gained currency in recent years for achieving that focus is broadly known as "process quality." Essentially, the idea is that the important thing is that we agree on how each performer of e-discovery services should design measures to gain insight into the quality of the results achieved by their particular process. The design of their process, and of their specific measures, is up to each performer. Of course, some performers might benefit from economies of scale by adopting measures designed by others, but because the measures must fit the process and because process innovation should be not just accommodated but encouraged, forced convergence on specific measures can be counterproductive. So process quality approaches seek to certify the way in which the measurement process is performed rather than what specifically is measured.

The Sedona Conference® recognized in 2009 that "the legal profession is at a crossroads," in terms of its willingness to embrace automated, analytical and statistical approaches to how law is practiced, at least in the area of e-discovery (The Sedona Conference, 2009). Sedona, EDRM and others have called upon industry and academia to assist in the development of standards, or at least best practices, in the area of IR. Against this backdrop, an e-discovery certification entity that could reduce the need for evidentiary hearings on the reasonableness of particular IR approaches would surely be welcome. What might such a certification entity look like? Although there is as of now no "right on point" model for such an entity, there are related standards and benchmarking entities to which we might look for inspiration.

Perhaps the best known example is the ISO 9000 family of international quality management system standards (International Organization for Standards, 2005). Standards in the ISO 9000 family serve two fundamental purposes: they provide guidance to organizations that wish to institute or improve their management of process quality, and they provide a basis for optionally certifying compliance by individual organizations with specific standards in the family. For example, a company or organization that has been independently audited and certified to be in conformance with ISO 9001:2008 (one standard in the family) may publicly state that it is "ISO 9001 certified." Such a certification does not guarantee quality; rather, it certifies that formal processes for measuring and controlling quality are being applied. These "quality management controls" include, for example, providing for internal audits and for taking corrective and preventive actions.

The basic ideas behind the ISO 9000 family of standards are also reflected in several other types of standards and best practice guidelines. For example, the Statement on Auditing Standards No. 70: Service Organizations (SAS 70) provides for audit reports that comment on the "fairness of the presentation of the service organization's description of controls[,] . . . the suitability of the design of the controls to achieve the specified control objectives[, and] ... whether the controls were operating effectively during the period under review" (American Institute of Certified Public Accountants, 2009) Similarly, the Payment Card Industry Data Security Standard (PCI DSS) prescribes controls for the processing of credit card payments and describes a process for certifying compliance with the standard (PCI Security Standards Council, 2009).

In software engineering, the Capability Maturity Model Integration (CMMI) plays a similar role, although with a less prescriptive approach. CMMI is a process improvement process that helps “integrate traditionally separate organizational functions, set process improvement goals and priorities, provide guidance for quality processes, and provide a point of reference for appraising current processes.”[40] Software development organizations can be appraised and given a maturity level rating or a capability level achievement profile.

At the present juncture, no such certification standard or body is on the horizon, but there is no theoretical barrier to their formation, provided that some consensus emerges in the legal community as to what types of certification would be appropriate. The true legacy of the present-day TREC Legal Track, the Sedona Search Commentary, the Electronic Discovery Institute study, and EDRM might therefore ultimately be their role as “midwife” to the creation of some (possibly international) standard-setting body.

7 Conclusion

Evaluation of IR for e-discovery is a complex task, but as we have seen substantial progress has already been made. We now have useful frameworks for thinking through the role of search technology, we have test collections with baseline results against which future progress can be measured, and we have seen the emergence of a robust research community with this important challenge as its focus. In the natural cycle between interactive and Cranfield-style evaluation, interactive evaluation presently occupies much of the spotlight. Cranfield-style evaluation has a good deal to say, however, and as the number of topics available for the new TREC Enron collection grows, further progress in Cranfield-style evaluation of e-discovery can reasonably be expected. Diversity is a hallmark of a maturing research community, so test collections for related tasks (e.g., de-duplication and privilege review) will also surely be developed. TREC is not likely to remain the principal evaluation venue for e-discovery indefinitely for the simple reason that TREC’s work in e-discovery will be done when sufficiently rich communities, collections, and comparative data (i.e., baseline results) are in place. But, with the narrow exception of evaluation design, TREC itself has never been the true center of gravity for e-discovery research. That honor goes to the research teams who have designed and built the systems and interactive processes that have been tried to date, and those that will be tried in the future.

Acknowledgements The authors first wish to thank a number of individuals who in discussions with the authors contributed ideas and suggestions that found their way into portions of the present paper, including Thomas Bookwalter, Gordon Cormack, Todd Elmer, Maura Grossman and Richard Mark Soley. Additionally, the TREC Legal Track would not have been possible without the support of Ellen Voorhees and Ian Soboroff of NIST; the faculty, staff and students of IIT, UCSF, Tobacco Documents Online, and Roswell Park Cancer Institute who helped build IIT CDIP or the LTDL on which it was based; Celia White (the 2006 Track expert interactive searcher); Venkat Rangan of Clearwell Systems who helped to build the TREC Enron test collection; Richard Braman of The Sedona Conference® and the hundreds of law students, lawyers and Sedona colleagues who have contributed *pro bono* time to the project. Finally, the authors wish to thank Kevin Ashley and Jack Conrad for their support

[40] Software Engineering Institute, “What is CMMI,” http://www.sei.cmu/edu/cmmi/general/index.html

of and participation in the First and Third DESI Workshops, held as part of the Eleventh and Twelfth International Conferences on Artificial Intelligence and Law, at which many of the ideas herein were discussed.

References

American Institute of Certified Public Accountants (2009) Statement on auditing standards no. 70: Service organizations. SAS 70

Aslam JA, Pavlu V, Yilmaz E (2006) A statistical method for system evaluation using incomplete judgments. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 541–548

Bales S, Wang P (2006) Consolidating user relevance criteria: A meta-ethnography of empirical studies. In: Proceedings of the 42nd Annual Meeting of the American Society for Information Science and Technology

Baron JR (2005) Toward a federal benchmarking standard for evaluation of information retrieval products used in e-discovery. Sedona Conference Journal 6:237–246

Baron JR (2007) The TREC legal track: Origins and reports from the first year. Sedona Conference Journal 8:237–246

Baron JR (2008) Towards a new jurisprudence of information retrieval: What constitutes a 'reasonable' search for digital evidence when using keywords? Digital Evidence and Electronic Signature Law Review 5:173–178

Baron JR (2009) E-discovery and the problem of asymmetric knowledge. Mercer Law Review 60:863

Baron JR, Thompson P (2007) The search problem posed by large heterogeneous data sets in litigation: Possible future approaches to research. In: Proceedings of the 11th International Conference on Artificial Intelligence and Law, pp 141–147

Baron JR, Lewis DD, Oard DW (2007) TREC-2006 Legal Track Overview. In: The Fifteenth Text REtrieval Conference Proceedings (TREC 2006), pp 79–98

Bauer RS, Brassil D, Hogan C, Taranto G, Brown JS (2009) Impedance matching of humans and machines in high-q information retrieval systems. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, pp 97–101

Blair D (2006) Wittgenstein, Language and Information: Back to the Rough Ground. Springer, New York

Blair D, Maron ME (1985) An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM 28(3):289–299

Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30(1-7):107–117

Buckley C, Voorhees EM (2004) Retrieval evaluation with incomplete information. In: Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval, pp 25–32

Buckley C, Voorhees EM (2005) Retrieval system evaluation. In: Voorhees EM, Harman DK (eds) TREC: Experiment and Evaluation in Information Retrieval, MIT Press, Cambridge, MA, pp 53–75

Buckley C, Dimmick D, Soboroff I, Voorhees E (2006) Bias and the limits of pooling. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 619–620

Büttcher S, Clarke CLA, Yeung PCK, Soboroff I (2007) Reliable information retrieval evaluation with incomplete and biased judgements. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 63–70

Carmel D, Yom-Tov E, Darlow A, Pelleg D (2006) What makes a query difficult? In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 390–397

Carterette B, Pavlu V, Kanoulas E, Aslam JA, Allan J (2008) Evaluation over thousands of queries. In: Proceedings of the 31st annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 651–658

Clarke C, Craswell N, Soboroff I (2005) The TREC terabyte retrieval track. SIGIR Forum 39(1):25–25

Cleverdon C (1967) The Cranfield tests on index language devices. Aslib Proceedings 19(6):173–194

Cormack GV, Lynam TR (2006) TREC 2005 spam track overview. In: Voorhees EM, Buckland LP (eds) The Fourteenth Text Retrieval Conference (TREC 2005), National Institute of Standards and Technology (NIST), vol Special Publication 500-266, pp 91–108

Cormack GV, Palmer CR, Clarke CLA (1998) Efficient construction of large test collections. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 282–289

Dumais ST, Belkin NJ (2005) The TREC interactive tracks: Putting the user into search. In: Voorhees EM, Harman DK (eds) TREC: Experiment and Evaluation in Information Retrieval, MIT Press, Cambridge, MA, pp 123–152

Fox EA (1983) Characterization of two new experimental collections in computer and information science containing textual and bibliographic concepts. Tech. rep., Cornell University

Harman DK (2005) The TREC text collections. In: Voorhees EM, Harman DK (eds) TREC: Experiment and Evaluation in Information Retrieval, MIT Press, Cambridge, MA, pp 21–52

Hedin B, Oard DW (2009) Replication and automation of expert judgments: information engineering in legal e-discovery. In: SMC'09: Proceedings of the 2009 IEEE international conference on Systems, Man and Cybernetics, pp 102–107

Hedin B, Tomlinson S, Baron JR, Oard DW (2010) Overview of the TREC 2009 Legal Track. In: The Eighteenth Text REtrieval Conference (TREC 2009), to appear

Ingwersen P (1992) Information Retrieval Interaction. Taylor Graham, London

Ingwersen P, Järvelin K (2005) The Turn: Integration of Information Seeking and Retrieval in Context. Springer

International Organization for Standards (2005) Quality management systems—fundamentals and vocabulary. ISO 9000:2005

Jensen JH (2000) Special issues involving electronic discovery. Kansas Journal of Law and Public Policy 9:425

Kando N, Mitamura T, Sakai T (2008) Introduction to the NTCIR-6 special issue. ACM Transactions on Asian Language Information Processing 7(2):1–3

Kazai G, Lalmas M, Fuhr N, Gövert N (2004) A report on the first year of the INitiative for the Evaluation of XML retrieval (INEX'02). Journal of the American Society for Information Science and Technology 55(6):551–556

Lewis D, Agam G, Argamon S, Frieder O, Grossman D, Heard J (2006) Building a test collection for complex document information processing. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 665–666

Lewis DD (1996) The TREC-4 filtering track. In: The Fourth Text Retrieval Conference (TREC-4), pp 165–180

Lynam TR, Cormack GV (2009) Multitext legal experiments at TREC 2008. In: Voorhees EM, Buckland LP (eds) The Sixteenth Text Retrieval Conference (TREC 2008), National Institute of Standards and Technology (NIST)

Majumder P, Mitra M, Pal D, Bandyopadhyay A, Maiti S, Mitra S, Sen A, Pal S (2008) Text collections for FIRE. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp 699–700

Moffat A, Zobel J (2008) Rank-biased precision for measurement of retrieval effectiveness. ACM Transactions on Information Systems 27(1)

Oard DW, Hedin B, Tomlinson S, Baron JR (2009) Overview of the TREC 2008 Legal Track. In: The Seventeenth Text REtrieval Conference (TREC 2008)

Oot P, Kershaw A, Roitblat HL (2010) Mandating reasonableness in a reasonable inquiry. Denver University Law Review 87:533

Paul GL, Baron JR (2007) Information inflation: Can the legal system adapt? Richmond Journal of Law and Technology 13(3)

PCI Security Standards Council (2009) Payment card industry (pci) data security standard: Requirements and security assessment procedures. URL http://www.pcisecuritystandards.org, version 1.2.1

Peters C, Braschler M (2001) European research letter: Cross-language system evaluation: The CLEF campaigns. Journal of the American Society for Information Science and Technology 52(12):1067–1072

Roitblat HL, Kershaw A, Oot P (2010) Document categorization in legal electronic discovery: Computer classification vs. manual review. Journal of the American Society for Information Science and Technology 61:70–80

Sakai T, Kando N (2008) On information retrieval metrics designed for evaluation with incomplete relevance assessments. Information Retrieval 11(5):447–470

Sanderson M, Joho H (2004) Forming test collections with no system pooling. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 33–40

Sanderson M, Zobel J (2005) Information retrieval system evaluation: effort, sensitivity, and reliability. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 162–169

Schmidt H, Butter K, Rider C (2002) Building digital tobacco document libraries at the University of California, San Francisco Library/Center for Knowledge Management. D-Lib Magazine 8(2)

Singhal A, Salton G, Buckley C (1995) Length Normalization in Degraded Text Collections. In: Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval, pp 15–17

Soboroff I (2007) A comparison of pooled and sampled relevance judgments. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 785–786

Solomon RD, Baron JR (2009) Bake offs, demos & kicking the tires: A practical litigator's brief guide to evaluating early case assessment software & search & review tools. URL http://www.kslaw.com/Library/publication/BakeOffs_Solomon.pdf

Spärck Jones K, van Rijsbergen CJ (1975) Report on the need for and provision of an ideal information retrieval test collection. Tech. Rep. 5266, Computer Laboratory, University of Cambridge, Cambridge (UK)

Taghva K, Borsack J, Condit A (1996) Effects of OCR errors on ranking and feedback using the vector space model. Information Processing & Management 32(3):317–327

The Sedona Conference (2007a) The Sedona Principles, Second Edition: Best practice recommendations and principles for addressing electronic document production. URL http://www.thesedonaconference.org

The Sedona Conference (2007b) The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery. The Sedona Conference Journal 8:189–223

The Sedona Conference (2009) The Sedona Conference Commentary on Achieving Quality in the E-Discovery Process. The Sedona Conference Journal 10:299–329

Tomlinson S (2007) Experiments with the negotiated Boolean queries of the TREC 2006 legal discovery track. In: The Fifteenth Text REtrieval Conference (TREC 2006)

Tomlinson S, Oard DW, Baron JR, Thompson P (2008) Overview of the TREC 2007 Legal Track. In: The Sixteenth Text Retrieval Conference (TREC 2007) Proceedings

Turpin A, Scholer F (2006) User performance versus precision measures for simple search tasks. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 11–18

Voorhees EM, Garofolo JS (2005) Retrieving noisy text. In: Voorhees EM, Harman DK (eds) TREC: Experiment and Evaluation in Information Retrieval, MIT Press, Cambridge, MA, pp 183–197

Voorhees EM, Harman DK (2005) The Text REtrieval Conference. In: Voorhees EM, Harman DK (eds) TREC: Experiment and Evaluation in Information Retrieval, MIT Press, Cambridge, MA, pp 3–19

Wayne CL (1998) Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In: Proceedings of the First International Conference on Language Resources and Evaluation

Yilmaz E, Aslam JA (2006) Estimating average precision with incomplete and imperfect judgments. In: Proceedings of the 15th International Conference on Information and Knowledge Management (CIKM), pp 102–111

Zhao FC, Oard DW, Baron JR (2009) Improving search effectiveness in the legal e-discovery process using relevance feedback. In: ICAIL 2009 DESI III Global E-Discovery/E-Disclosure Workshop, URL http://www.law.pitt.edu/DESI3_Workshop/DESI_III_papers.htm

Zobel J (1998) How reliable are the results of large-scale information retrieval experiments? In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 307–314