Improving Search Effectiveness in the Legal E-Discovery Process Using Relevance Feedback

Feng C. Zhao
School of Law
University of Washington
fcz@u.washington.edu

Douglas W. Oard
Coll. of Info. Stu. & UMIACS
University of Maryland
College Park, MD 20742
oard@umd.edu

Jason R. Baron
Office of the General Counsel
National Archives and
Records Administration
jason.baron@nara.gov

ABSTRACT

Finding information and preserving confidentiality are in some sense opposite goals, but there are a number of practical situations in which a balance must be struck between the two. Identifying relevant evidence in large collections of digital business records, the so-called "e-discovery" problem, is one such situation. Present practice involves consultation between attorneys representing plaintiffs and defendants, a search conducted by defendants (or their agents) in response to plaintiffs' requests, manual review of every retrieved document, and release to the plaintiffs of all documents that are judged relevant and not subject to a claim of privilege.

Although advanced search strategies and tools are available, the keyword based search dominates current legal practice in e-discovery as it is well understood and has been commonly used by the legal community for a long time. However, it is difficult for a party to select the right keywords to achieve a satisfying recall level without knowledge of the other party's data.

This paper applies relevance feedback and result fusion techniques as a simple model for a multi-stage consultation process. Experiments using the TREC Robust Track test collection and relevance feedback using offer weights show that a partial release of relevant documents, followed by a second consultation, has the potential to substantially improve overall retrieval effectiveness, and that additional partial releases and subsequent consultations seem to offer diminishing potential for additional benefit.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query formulation, Relevance feedback, Search process

General Terms

Design, Experimentation, Measurement, Performance

Keywords

Legal E-Discovery, Information Retrieval, Relevance Feedback, Query Expansion

1. INTRODUCTION

The legal profession and their 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.[2] The phenomenon known as "e-discovery," i.e., the requirement that documents and information in electronic form stored in corporate databases and networks be produced as evidence in litigation, is both 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 some form of sanctions to lawyers and their clients.[3] In turn, a spotlight has now formed on *how* lawyers decide to meet their obligations in various e-discovery contexts, one major aspect of which 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 do so. The "how" is increasingly important, given spiraling legal costs, with just one example being a Forrester report issued in 2006 estimating that the secondary market for information technology solutions in the legal tech sector will grow just in the U.S. to $4.8 billion by 2011.[4]

[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 1234 (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 generally, The Sedona Principles, Second Edition: Best Practice Recommendations and Principles for Addressing Electronic Document Production (2007), available at www.thesedonaconference.org.

[3]See, e.g., Qualcomm v. Broadcom, 539 F.Supp.2d 1214 (S.D. Cal 2007).

[4]The Forrester study is cited at http://www.forrester.com/Research/Document/Excerpt/0,7211,40619,00.html.

It nevertheless may strike some as incredible that even well into the present decade lawyers in the most complex litigation failed to employ any particularly advanced strategies for developing search protocols as an aid in conducting either manual and automated searches for legal information – 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 new federal rules of civil procedure, applicable to all federal courts and now increasingly being incorporated in many state courts. The new rules expressly reference the term "electronically stored information" (otherwise known as "ESI"), that impose new requirements on lawyers at the initial stages of litigation to engage in a robust "meet and confer" process – where issues going to the location, preservation, and formatting of, as well as access to an adversary party's ESI are to be discussed.[5] These new requirements serve to in turn spotlight a real gap in legal practice, in the legal profession's difficulty identifying and adopting best practices and proven techniques from the well-studied field of information retrieval. Our goal in this paper is to contribute to that discussion by using well understood techniques for query reformulation and for evaluation to explore the potential of multi-stage negotiations to improve search effectiveness.

2. E-DISCOVERY SEARCH MODELS

In this section we describe the evolution of thinking about supporting e-discovery using search technology.

2.1 Legal Discovery before the "E-"

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, have 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 stated named topic - were propounded with the expectation that the receiving party would perform a reasonably diligent search for records found in corporate hard-copy 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; 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: crafting requests in the broadest conceivable way; the receiving 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 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 hard copy 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.

2.2 The State of the Art: "Keyword Search"

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, changed the legal terrain considerably. Lawyers familiar with structured databases of cases and legislation, as created by Lexis and Westlaw, became readily adept at using keywords 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 unstructured corporate databases. As one court put it, "[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."[7] Most often, this involves simply searching for a single term (i.e., a single word or phrase). As recent legal scholarship has shown, however, simple term matching suffers from a variety of known 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, i.e., the 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," throughout the discovery

[5]Jones v. Goord, 2002 WL 1007614, *1 (S.D.N.Y. May 16, 2002).

[6]See U.S. Federal Rules of Civil Procedure, amended Dec. 1, 2006; see generally, The Sedona Principles, Second Edition, supra, 2.

[7]In re Lorazepam & Clorazepate Antitrust Litigation, 300 F. Supp. 2d 43, 46 (D.D.C. 2004).

[8]See The Sedona Conferenceő Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery, 8 Sedona Conf. J. 189 (2007), available at http://www.thesedonaconference.org/

process.[9]

In the past few years, and especially with the advent of the 2006 federal rules of civil procedure in the U.S., change has been in the air: a variety of published decisions have been handed down recognizing the need for parties to undertake some form of limited collaboration with respect to search protocols, most notably on the issue of exchanging proposed "keywords" for use in possible searches.[10] To an even more limited extent, a few courts have recognized that there are more sophisticated means of employing search strategies, from using Boolean operators,[11] to considering alternatives to Boolean searches in the form of conducting what has been deemed "concept searching,"[12] to using other forms of search models incorporating probability, Bayesian, and/or fuzzy elements.[13] One court has gone so far to suggest that given the interplay of statistics and "linguistics" in the matter of constructing search queries, that lawyers may need to introduce forms of "expert" testimony as an aid to the fact finder.[14] Although the "jury is still out" on the proven efficacy of using alternative, non-Boolean forms of searching [3], there can be little doubt that the legal profession (and corporate and institutional clients of all stripes) can collectively benefit from demonstrated improvements in the science of information retrieval, from both the perspective of accuracy (i.e., finding would-be buried evidence), and efficiency (i.e., reducing costs). Even just leveraging from some common information retrieval technologies, we can still improve the current e-discovery process, with existing keyword based search platform, to achieve a new level of retrieval effectiveness.

2.3 Iterative Query Reformulation

Importantly, for purposes of the present paper, in the vast majority of legal settings, there is 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 knows its contents). The extent to which asymmetry in information sharing between adversaries in litigation is a problem depends to a large extent on the nature of the information retrieval task being conducted.

Specifically with respect to the type of "task" to be performed, three types of searches present themselves typically in litigation:

Type I: Known-item search: users are looking for specific items whose characteristics are known and can be used for search. For example, plaintiffs request the disclosure of a file with the exact file name based on prior knowledge that the requested document contains the particular trade secret at issue. As the result, only the targeted item previously known to the searcher is relevant.

Type II: Known-topic search: users are interested in items that address certain topics and have at least a conceptual model of how to confine a search to those topics. Most e-discovery practice today falls into this category. In this case, the relevance of the search results ultimately a matter of judgment for domain experts (or, when collection sizes exceed what domain experts could possibly review, for large review teams in which individual reviewers may have no special domain expertise).

Type III: Exploratory search: users simply want to explore available resources for potential evidence, without any predefined forms. Much like the fictional detective Sherlock Holmes inspecting the crime scene, lawyers may wish to systematically or randomly scan through ESI repositories without predefined search targets. Certainly, Type III exploration searches can evolve into Type II known topics search, and often may be accompanied by dynamically changing topics. Relevance is arguably a more ephemeral concept here that evolves with the search topic(s).

The above three types of searches can well co-exist in a given e-discovery process. In a recent a trade secret case,[15] the plaintiff wished to obtain images of defendant's servers without any limitation in conducting the search. The defendant in turn wanted to protect its confidential information, so as to limit the search to the files that might contain trade secret information only, such as files with unique electronic fingerprints (or MD5 hash values), identified file names, or certain keywords provided by the plaintiff. File name matching and MD5 hash value matching both belong to Type I known items searches. The keyword-based search here would be a Type II known topics search. The plaintiff's proposed search falls into the Type III exploration search due to its undefined scope.

Our focus in this paper is on the Type II known-topic search. From the information requester's perspective, there is a wide and varying spectrum of possibilities representing the requestor's "state of knowledge" in a given e-discovery setting of the opposing parties' data store. For illustrative purposes, we choose the following simple model representing three such states:

Blind: when the requesting party lacks knowledge of the responding party's information system and data, the e-discovery requests are based on general knowledge and common sense, supplemented (in the most sophisticated forms of litigation) by whatever background information can be acquired through the use of an interdisciplinary team of experts. After a "one-time" meet-and-confer, opposing parties may agree on a set of keywords, or perhaps on a more sophisticated query involving Boolean and proximity operators, for search —that effectively becomes the basic search protocol. The responding party in turn fulfills its obligation by producing a set of relevant items. Accordingly, regardless whether the information request appears initially too broad or too narrow, what you see is what you get. E-discovery

[9]See Jason R. Baron, "E-Discovery and the Problem of Asymmetric Knowledge," 60 Mercer L. Review 863 (2009).

[10]See, e.g., William A. Gross Construction Associates, Inc. v. Am Mftrs. Mutual Ins. Co., 2009 WL 724954 (S.D. N.Y. March 19, 2009); Spieker v. Quest Cherokee, 2008 WL 4758604 (D. Kan. Oct 30, 2008);Treppel v. Biovail, 233 F.R.D. 363 (S.D.N.Y. 2006). We note that at least one court in the United Kingdom similarly has analyzed keyword choices by parties at some length. See Digicel (St. Lucia) Ltd. & Ors. v. Cable & Wireless & Ors., [2008] EWHC 2522 (Ch.).

[11]See ClearOne Communications, Inc. v. Chiang, 2008 WL 920336 (D. Utah, April 1, 2008); Williams v. Taser Intern, Inc., 2007 WL 1630875 (N.D. Ga.).

[12]See Disability Rights Council of Greater Washington, et al. v. Washington Metropolitan Transit Authority, 242 F.R.D. 139 (D.D.C. 2007).

[13]See Victor Stanley, Inc. v. Creative Pipe, Inc., 2008 WL 2221841 (D. Md.).

[14]See United States v. O'Keefe, 537 F. Supp. 2d 14 (D.D.C. 2008) .

[15]Bro-Tech Corp. v. Thermax, Inc., 2008 WL 724627 (E.D. Pa. March 17, 2008).

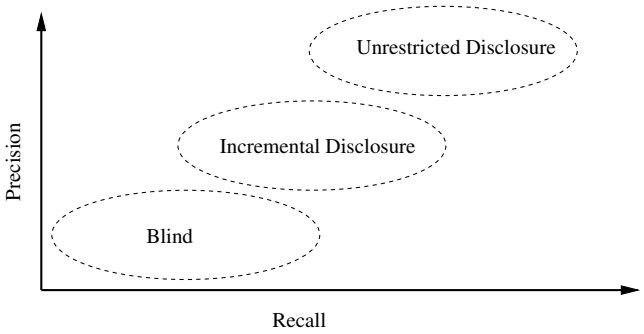

Figure 1: E-Discovery States

is a linear process in this setting and search is generally a one-time "take it or leave it" deal.

Incremental disclosure: when the requesting party is allowed to interact with the responding party's information system (either under a court-supervised, multiple-stage meet and confer process, or in a form of feedback loop through repeated communications between opposing counsel and/or more directly with the actual system and/or IT staff), e-discovery becomes an iterative and escalating process. The requesting party can leverage the system response to reformulate and refine the queries. Therefore, the goal for the requesting party is to formulate the optimal query and achieve the highest recall level possible. Note, however, that given sometimes limited willingness of the responding party as well as limitations in information system, the full potential of the query refinement might not be realized.

Unrestricted disclosure: when the requesting party gains full access to the responding party's information system and data, then the whole corps of information retrieval tools can be mobilized within the limitations of available resources. This may occur when a court has ruled that cost-shifting is appropriate and that one party (such as the requestor) should bear the costs of a request for information, including the initial obligation to search for relevant evidence.

In the current e-discovery landscape, the blind setting is still the overwhelmingly dominant form. However, even with a tremendous advantage one side has in knowing its own data set, in the vast majority of settings neither party takes advantage of knowing what results will be obtained. For example, it is not presently even common to determine the number of documents that would be found in the results set from a negotiated query. This lack of knowledge forms a baseline; it represents an "as is" model as currently practiced by the legal profession that arguably is open to improvement if information is gained—and shared—regarding the results obtained from initial searches. Our focus in this paper is therefore on exploring the incremental disclosure setting. In particular, we are interested in whether this can improve recall, and if so what the optimum number of rounds of renegotiation would be and where the optimum timing for those renegotiation points would be. While there have been some calls in the past couple of years for lawyers to engage in an iterative process of negotiations [4], we are not aware of any cases in which courts have analyzed or reported on the results of such negotiations.

2.4 Information Retrieval Experiments

Two prior research projects provide a useful framework for our experimental design notion here for improving search effectiveness in the e-discovery process.

The seminal study evaluating the success of text retrieval methods in an e-discovery setting was conducted by Blair & Maron over 20 years ago [2]. That study identified a serious gap between the perception on the part of lawyers that using keywords 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 on the basis of using additional keywords that only 20% of relevant documents had in fact been found. These results have been replicated in other settings, and from them we can conclude that estimating recall is simply a hard task for people to do. We therefore want to avoid study designs in which accurate estimates of recall would be required.

As is often the case, this seminal work was in some sense before its time, since large collections of records in digital form are a considerably more recent development. In 2006, a new "Legal Track" was started as part of the Text REtrieval Conference (TREC)[16] (which is run by the U.S. National Institute of Standards and Technology) as one way of beginning to draw the information retrieval and e-discovery communities more closely together. The basic protocol of the legal track's "ad hoc" task has been to engage in (a) constructing a series of hypothetical complaints; (b) constructing a large variety of what are known as "requests to produce documents" which for simplicity we will refer to here as a "topic;" and (c) negotiating between two lawyers a search protocol for each topic using Boolean, proximity and truncation operators should be used to conduct searches. The negotiations carried out as part of the TREC Legal Track topic broke down into a three-stage query negotiation, consisting of (i) an initial query, as proposed by the receiving party on a discovery request, usually reading the request narrowly; (ii) a "counter"-proposal by the propounding party, usually including a broader set of terms; and (iii) a final 'negotiated' query, representing what was deemed the consensus arrangement as agreed to by the parties without resort to further judicial intervention. A team of volunteer lawyers and law students have thereafter evaluated sampled results for the purpose of assessing binary relevance: either a given document found by one or more search methodologies is relevant to the topic, or it isn't. In this fashion, nine complaints, 86 topics, and ultimately 56,142 relevance judgments have been generated in the first two years of the research [1, 7].

One of the clearest results from the TREC legal track is that many relevant documents are missed by the best present search methods [1, 7]. For example, in 2007, the second year of the track, the negotiated queries identified only 22% of the total known relevant documents—that left at least 78% to be found by some other means. Of course, the alternative systems that found those relevant documents could take advantage of ranking algorithms, automated approaches to query expansion (e.g., blind relevance feedback), stemming, and other advanced techniques, so the limited effectiveness of an exact-match query built using a blind process (i.e., a process in which nobody looked at any results during the negations) should not be surprising. What is

[16]The Text REtrieval Conference (TREC), available at http://trec.nist.gov.

surprising, however, is that no other single system found a greater number of relevant documents (on average, across many topics) than the negotiated query (when limited for purpose of evaluation to the number of documents returned by the negotiated query).

Although the extent to which the parties were aware of what they were in fact accomplishing when conducting their Boolean negotiations is not known, some of the resulting effects can be measured. Zhao, et al. [10] concluded that "[t]he primary techniques used in the Boolean refinement process are enriching the query with synonym-like terms and relaxing Boolean constraints" [10]. That paper went on to note a basic infertility of Boolean query negotiations, as practiced, given that "the negotiation[s] [were] ineffective to discover and inject semantically independent terms into queries. In other words, the negotiated final query essentially has the same semantic coverage as what was initially proposed by the defendant after we drop all of the Boolean syntax." In a similar vein, Tomlinson found that in the initial Boolean proposal by the receiving party (i.e., the party with an interest in limiting the scope of its own production obligations) reliably always recovered fewer documents than the final negotiated query [7]. Conversely, the rejoinder proposal by the propounding party of the discovery (designated plaintiff in the legal track), with only limited exceptions, resulted in a much larger set of retrieved documents than the set finally negotiated by the parties.

These results from fully automated experiments have motivated several preliminary experiments with iterative refinement. For example, in 2006 a domain expert used an interactive system to identify 100 documents per topic with the principal goal of enriching the set of known relevant documents. When scored using R-precision, that domain expert outperformed every automated system (regardless of how many documents the automated system returned) [1]. This led to creation of a new "interactive task" in 2007 in which eight teams explored alternative approaches to interactive query reformulation [7]. User studies of this sort offer significant potential for inspiration, but reliable measurement of specific effects would incur tremendous costs because human nature adds significant uncontrolled variability that can only be controlled for when very large numbers of users are employed. Moreover, the results from user studies are not easily reused to test alternative hypotheses that were not included in the initial study design.

As an alternative to user studies, a pilot "relevance feedback" task was also tried in 2007. In that task, 10 topics from 2006 for which relevance judgments were available were made available to participating teams for use as a basis for identifying additional terms that could be productively added to a query (or for any other fully automated purpose). Although some improvement was demonstrated from the use of relevance feedback, perhaps the most important conclusion from these initial experiments was that the characteristics TREC legal track test collection add a good deal of complexity to the design of relevance feedback experiments. The reason for this is that the collection consists entirely of scanned documents for which the content has been recovered using Optical Character Recognition (OCR), and the OCR accuracy varies markedly (e.g., because of degraded images and because some documents consist entirely of handwriting). Some metadata describing each document is also available, and jointly using OCR of variable quality along with structured metadata is indeed an interesting research problem. But it is not the problem in which we are interested here. We therefore have chosen to use a "born digital" collection of electronic text for the experiments that we report in this paper.

3. EXPERIMENT DESIGN

With that as background, we can now begin to design a set of experiments in this section.

3.1 Making Choices

Our goal is to inform the current practice of e-discovery, keyword based search, by exploring alternatives to the present practice of what is essentially a blind discharge of the meet-and-confer obligation. In particular, we have chosen to focus on an incremental disclosure approach to what we have called Type II known-topic search. That focus leads to the following research questions:

RQ1 Can incremental disclosure be used with query renegotiation to increase the number of relevant documents found without increasing the total manual review workload?

RQ2 How many rounds of query renegotiation are needed before a point of diminishing returns is reached?

RQ3 By what criteria should decisions be made about when each incremental disclosure and renegotiation should be conducted?

Attempting to address such a broad sweep of research questions requires that we adopt an affordable and repeatable methodology. One reasonable approach in such cases is to leverage an information retrieval test collection by using a fixed set of relevance judgments to score multiple variants of a process model. Such an approach requires compromises, of course. In particular, the process must be fully automated if a large number of variants are to be explored, the conclusions will be most useful if the topics, the documents, and the evaluations measure(s) are representative of the envisioned application, and the evaluation measure(s) must be reliably computable for all system variants using the existing relevance judgments. While none of these desiderata can be achieved perfectly, we can approximate each. In particular:

- We model the query renegotiation using well known techniques for "relevance feedback." the automatic enrichment of an existing query based on the statistics of term occurrence in documents that are known to be relevant. Existing approaches to relevance feedback models can learn not just presence but also term weights. Learning Boolean, proximity and truncation operators would also likely be feasible (e.g., using rule induction), but we have chosen to focus on simpler models that learn only term weights in order to avoid introducing additional complexity at this stage in our work. Our initial queries are therefore simple term lists, and our results are enriched ranked term lists.

- We use an information retrieval test collection from the TREC Robust Track. The topics in this collection were selected from a larger universe of TREC topics with a focus on those for which one-pass (i.e., blind)

query formulation had proven in earlier evaluations to yield relatively poor results. This comports well with our goals—if there is a benefit to be obtained from incremental disclosure, these are exactly the types of topics where we would expect that benefit to be most apparent. A simple term-list query is provided with the collection; this matches our experiment design well.

- We model incremental disclosure by progressing down the ranked list from the top and establishing a decision rule that fires when a query update using relevance feedback should be performed. The documents in the list above that point are frozen in place so that the subsequent refined query can affect only the remainder of the ranked list. We have explored two broad classes of decision rules, one based on the number of documents that have so far been reviewed, and a second based on the number of relevant documents that have so far been found (which can be at least approximately known during the review process).

- As an evaluation measure, we use the total number of relevant documents found with a fixed level of effort. We measure this as recall at a fixed cutoff (e.g., 10,000 documents). Because we are more interested in expected performance on future topics than in the results for particular pre-defined topics, we compute the expected value of this recall measure over all topics. This computation is known to be problematic for small cutoffs (e.g., 10 documents) in experiments modeling interactive use because some topics will have fewer relevant documents than the cutoff and others will have many more. We model the e-discovery with far larger cutoffs, however, so stability of this measure is less of a concern in our case. Note that we are focused here only on review for relevance; review for claims of privilege is not modeled in our experiments.

- Any set of relevance judgments for a large information retrieval test collection will naturally be both incomplete (because sampling strategies are needed to constrain the cost of constructing the collection) and somewhat eclectic (because opinions regarding relevance vary to some degree between individuals). Earlier experiments have, however, confirmed that substantial relative differences in effectiveness measures are nonetheless usually robust indicators of true differences when comparing fully automated systems [8]. We therefore focus principally on relative differences in our analysis. Reporting only differences in the mean can, however, mask significant variability that would be important in operational settings. We therefore report statistical significance using a two-sided t-test for paired samples (at $p < 0.05$) when comparing specific pairs of mean values.

3.2 Computational Models

In this section, we formalize our computational approach to modeling query renegotiation based on e-discovery.

3.2.1 Test Collection

We chose the TREC 2005 Robust Track test collection [9], which was in turn built from the AQUAINT collection. The AQUAINT collection consists of about a million English news stories from the New York Times, the Associated Press, and the Xinhua News Agency. For each of the 50 topics in the collection, we used the words in the title field of the topic statement as the initial query. The name of the "title" field is somewhat anachronistic in TREC usage; the words in this field are generally intended to model the initial (i.e., "blind") query that an interactive searcher might pose. Potential additional query terms are available in other fields of the topic statement, but experience has shown that so-called "title queries" often are as good as longer queries when used with ranked retrieval systems. We therefore adopted these title queries as a suitable starting point for our relevance feedback experiments.

3.2.2 Relevance Feedback

When renegotiating the query, both the plaintiff and the defendant will have access to the documents in the initial (partial) review set that are known to be relevant, and the defendant will additionally have access to the documents that are known not to be relevant. One natural goal in such cases is to find an improved query that will tend to find more documents that are similar to those in the relevant set, and dissimilar from those in the set of documents that are known by the defendant not to be relevant. This is modeled by adding some fixed number of terms that are selected based on their offer weights [6] as shown in Equation 2.

$$RW = r_i \log \frac{(r_i + 0.5)(N - n_i - R + r_i + 0.5)}{(n_i - r_i + 0.5)(R - r_i + 0.5)} \quad (1)$$

$$OW = r_i \, RW \quad (2)$$

RW	relevance weight
OW	offer weight
N	the collection size
R	number of relevant documents for a topic
n_i	number of documents in which term i occurs
r_i	number of relevant documents in which term i occurs

The prior query offers some additional evidences for what is sought that may not be fully captured by the terms selected using offer weights. As is common practice, we use a simple linear combination as introduced by Rocchio to combine the new and existing query term weights as shown in Equations 3 and 4. We empirically set $\beta = 0.5$ in our experiments.

$$OTW = \log \frac{N}{n_i} \quad (3)$$

$$ETW = \beta \, RW_y \frac{Y \sum^X OTW_x}{X \sum^Y RW_y} \quad (4)$$

OTW	weights for original query terms
ETW	weights for expanded query terms
RW_y	relevance weight for expanded query term y
X	number of original query terms
Y	number of expanded query terms
β	smoothing factor

3.2.3 Fixed-Partition Decision Rule

Let N_t be the number of documents to be reviewed for topic t; in our experiments we set N_t to the same value

$\forall t$. The plaintiff and defendant can then agree on any way of selecting a partition P_n smaller than N_t for relevance feedback. In this setting, the number of documents to be reviewed in each relevance feedback stage is fixed, while the number of relevant documents found during that stage is variable. In some initial experiments, we tried some simple models for fixed partition sizes:

$$P_n = a + N_t/k \quad \text{arithmetic progression} \tag{5}$$

$$P_n = a\, r^{n-1} \quad \text{geometric progression} \tag{6}$$

n	iteration sequence number
t	topic number
k	a positive integer number
a	a starting value
r	a ratio
P_n	partition size at iteration of n
N_t	upper bound of documents to be reviewed for topic t

3.2.4 Variable-Partition Decision Rule

A potential drawback of a fixed partition for the requesting party is that the density of relevant documents is not known in advance, and thus a partition that is adequate for one topic might be too small to contain enough relevant documents for some other topic. A natural alternative is to model the responding party as continuing their review until some fixed number of relevant documents have been found. A maximum partition size is, however, still needed to accommodate topics with exceptionally low densities of relevant document densities,

3.2.5 Evaluation Measure

As an evaluation measure, we have selected recall at N_t (i.e., the fraction of the known relevant documents that are at or above rank N_t in the final ranked list). For incremental disclosure, the position of all reviewed documents is frozen upon disclosure. This is one of several commonly used evaluation designs for relevance feedback experiments, and the one that best matches our goal of producing results that can be directly compared across multiple conditions while faithfully modeling a fixed review effort constraint. Of course, some effort must also be devoted to renegotiate the query at each iteration, so ultimately the evaluation comes down to a tradeoff between cost (of additional incremental disclosure stages) and benefit (as modeled by recall at N_t).

4. RESULTS

Our first set of experiments were designed to explore alternative designs for fixed-partition decision rules. For the first experiment, we set $N_t = 1K$ (i.e., we truncated each ranked list at 1,000) and we partitioned the resulting set into equal-sized regions using an arithmetic progression according to Equation 5. Table 1 shows the resulting partitions. For example, the "500-500" entry for $k = 2$ means that for 2 partitions the available relevance judgments for the first 500 documents will be used as a basis for relevance feedback, and then another 500 documents (after freezing the first 500) will be retrieved using the improved query.

This also provided an opportunity to tune the number of expansion terms that are to be added at each iteration. As Figure 2 shows, 5 expansion terms is not enough, 10 might be a reasonable choice, and 15 is plenty. Since adding query terms increases the query processing time, we have limited the number of expansion terms per iteration to 15 consistently throughout our remaining experiments.

k	Progression
1	1000
2	500-500
3	333-333-334
4	250-250-250-250
5	200-200-200-200-200
6	166-166-166-166-166-170
7	142-142-142-142-142-142-148
8	125-125-125-125-125-125-125-125
9	111-111-111-111-111-111-111-111-112
10	100-100-100-100-100-100-100-100-100-100

Table 1: Arithmetic progression partitions.

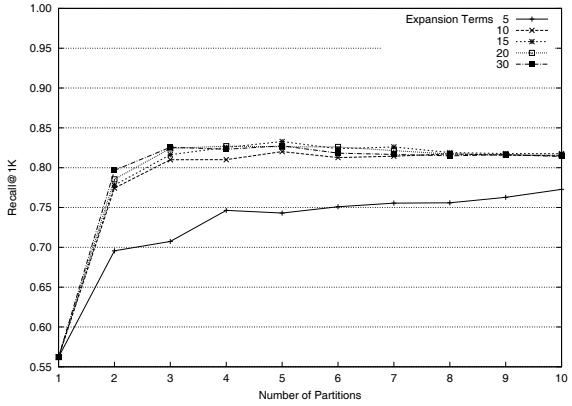

Figure 2: Recall@1K for arithmetic progression.

As Figure 2 shows, relevance feedback is a big win for this collection regardless of the partitioning details. This is not a surprise, of course—relevance feedback has repeatedly been shown to be helpful in fully automated experiments such as these. What is interesting, however, is that as Table 2 shows, a paired t-test reveals that performing relevance feedback a second time (i.e., at $k = 3$) results in a further statistically significant improvement in recall. The third and fourth iterations of relevance also seem as if they may do a bit of good, although the improvements are small enough at that point that the added benefit might not justify the added cost of renegotiation in actual e-discovery settings. We must caveat this result, however, by observing that our experiment design varies both the number of partitions and the number of documents in a partition simultaneously. In other words, perhaps 666 documents is better than 500 as a basis for relevance feedback, regardless of whether they are used in one or two iterations. We will come back to this point below.

As Table 3 shows, our geometric progression in Equation 6 results in a very small initial partition followed by progressively larger subsequent partitions. For these experiments we set the starting value $\alpha = 1$ and let the ratio vary $r \in [1 \ldots 10]$. The intuition behind these choices is

Pairs	p	Relative Change
1-to-2	**<0.001**	+38.4%
2-to-3	**<0.01**	+4.9%
3-to-4	0.18	+1.1%
4-to-5	**<0.05**	+0.9%
5-to-6	**<0.05**	-1.1%

Table 2: Changes in Recall@1K for arithmetic progression (bold=significant).

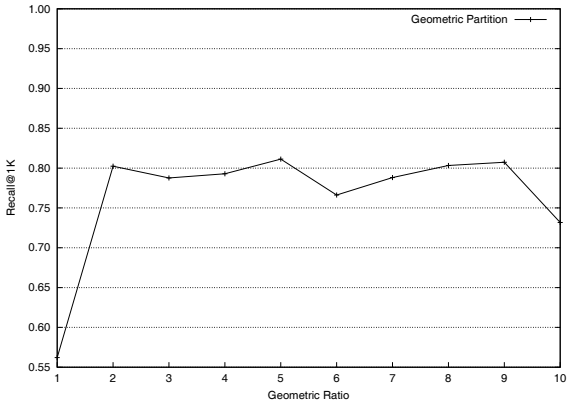

Figure 3: Recall@1K for geometric progression.

that rapidly leveraging early learning might pay dividends. But if that is the case, we can't see evidence for it in Figure 3, which plots the results for each value of r (as always henceforth, with 15 feedback terms). No variant of geometric progression that we tried outperforms our best results from the arithmetic progression.

r	Progression
1	1000
2	2-4-8-16-32-64-128-256-490
3	3-9-27-81-243-637
4	4-16-64-256-660
5	5-25-125-625-220
6	6-36-216-742
7	7-49-343-601
8	8-64-512-416
9	9-81-729-181
10	10-100-890

Table 3: Geometric progression partitions.

Our experiments with fixed-partition decision rules proved to be useful for tuning the number of expansion terms to use, for confirming that relevance feedback can be useful for this collection, and for illustrating that choices that we make about when to enhance the query using relevance feedback do matter. But both our arithmetic progression and our geometric progression had the undesirable characteristic that we varied the number of partitions and the size of those partitions at the same time. For our remaining experiments we therefore decided to measure retrieval effectiveness in a finer grained way, computing the recall achieved by each query in a sequence. Because Figure 2 suggests that a substantial benefit can be achieved with just a few iterations, we elected to focus on plots in which the number of relevant documents in the first iteration is set and then the number of relevant documents in the second iteration is allowed to vary. Figure 4 illustrates this idea, which we explain in more detail below.

A second factor that complicated the interpretation of our fixed-partition decision rule results is that the optimal number of documents to review might vary with the nature of the topic. It is well known that some topics are much harder than others, and even among the topics selected for the TREC 2005 Robust track we would expect this to be true. An initial 500-document partition, for example, might therefore be too large in some cases, and too small in others. In information retrieval experiments it is usually not considered fair to base system decisions on actual relevance judgments (since if you had the judgments you would not need the system!), but the e-discovery setting is different because (at least at present) the normal practice is to manually review documents prior to their disclosure. This key difference makes possible protocols in which the decisions about incremental disclosure are based on the number of relevant documents found so far, rather than simply on the number of documents reviewed so far. This is the basis for what we call a variable-partition decision rule: the total number of documents in a partition is set in a way that is designed to place some desired number of relevant documents in each partition. Of course, it is possible that the desired number of relevant documents may not actually exist, so we also need to bound the maximum size of a partition to avoid the case in which the first partition expands to N_t, thus preventing any relevance feedback at all.

Now, let us compare recall levels for a variable-partition decision rule across a greater range of evaluation cutoffs. The results in the "25|500" column of Table 4 were obtained with one iteration of relevance feedback after finding the first 25 relevant documents for each topic (what we call the "target"), or the first 500 total documents (what we call the "limit"), whichever came first. This approach clearly works at least as well as our best fixed-partition experiments for Recall@1K. Importantly, substantial improvements are also evident even when far more documents are selected for manual review, as this is becoming increasingly common in complex litigation. As the final column shows, a larger limit of at most 2,000 total documents in the first partition yields some further improvement in recall, but only for evaluation cutoffs larger than 2,000. The lack of benefit for evaluation cutoffs of 2,000 and below occurs because 9 of the 50 topics miss the target and hit the limit; for those topics, the first partition grows to encompass the entire evaluated set, thus precluding any benefit from relevance feedback. For the remainder of our experiments, we therefore focus on evaluation cutoffs of 5,000 and 10,000 (which we believe to be reasonable values for the number of documents that might require review in today's increasingly complex litigation) with a partition size limit of 2,000 documents.

Figures 4 and 5 show the effects of adding a second stage of relevance feedback to these variable-partition decision rules. Each line represents a specific target number of relevant documents for the first partition, with values along the X axis showing alternative targets for the second partition. For example, the lowest line in Figure 4 (i.e., the solid line in that

N	R@N no RF	R@N 25|500	Relative Change	p	R@N 25|2000
1K	0.5621	0.7871	+48.2%	**<0.001**	0.7225
2K	0.6666	0.8784	+33.0%	**<0.001**	0.8199
5K	0.7776	0.9351	+23.6%	**<0.001**	0.9481
10K	0.8497	0.9596	+16.2%	**<0.001**	0.9711

Table 4: Relative improvements from relevance feedback with a variable partition decision rule (bold=significant).

figure) shows that with a target of 5 relevant documents in the first partition and a target of an additional (different) 5 relevant documents in the second partition, a recall of 0.93 is achieved. The leftmost value for each line (at $X = 0$) shows the results of performing only a single instance of relevance feedback. For example, the next (dashed) line up shows that for a first target of 10 relevant documents and no second partition a recall of 0.94 is achieved. Somewhat counterintuitively, this indicates that if a total of 10 relevant documents will be used for relevance feedback, it does not help (and may actually hurt!) to break them up into two sets of five and do two stages. This suggests that 5 documents may simply not be enough to reliably obtain an improved query using our fully automatic method. Indeed, the same pattern is evident for two rounds of 10 relevant documents each or one round of 20 relevant documents—larger initial sets are clearly helpful, at least up to a point. The same pattern is evident in Figure 5.

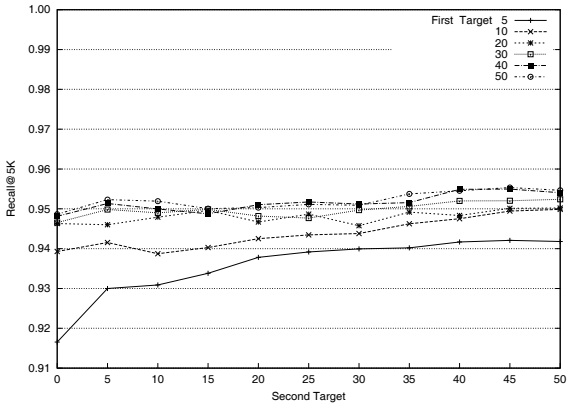

Figure 4: Recall@5K for a variable partition decision rule.

From inspection of Figures 4 and 5, it seems that the sweet spot is located somewhere around 30 relevant documents in the first partition for this test collection. That is, on average, about one quarter of the relevant documents for a topic, and we might reasonably expect this number to vary a bit from one collection to another (and certainly from one topic to another!). Moreover, it is important to point out that this is the number of examples of relevant documents that our automated algorithm needs in the first partition; experienced lawyers might achieve good results from fewer examples.

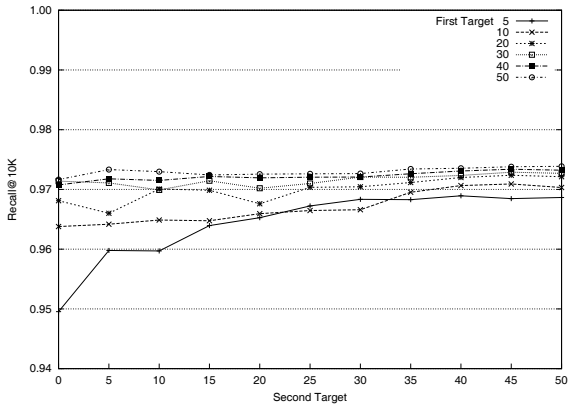

Figure 5: Recall@10K for a variable partition decision rule.

The more surprising result, however, is that if we set the first partition size in this way (a target of 30 relevant documents, with a limit of 2,000 total documents), no benefit is evident from *any* second stage of relevance feedback.

This set of experiments effectively answers our three research questions. To the extent that our experiments reasonably model actual human behavior, we have shown that:

RQ1 incremental disclosure can be used with query renegotiation to increase the number of relevant documents found without increasing the total manual review workload.

RQ2 One round of query renegotiation seems sufficient.

RQ3 A variable-partition decision rule in which a relatively small number of relevant documents (e.g., two to three dozen) are used as a basis for the renegotiation seems to work fairly well, but with the caveat the partition now be allowed to grow larger than some reasonable limit (e.g., half the total number of documents to be reviewed).

That's not to say, however, that no benefit could possibly accrue from multi-stage relevance feedback. Consider, for example, the results shown in the last line of Table 5, which were obtained by performing relevance feedback after *every* relevant document was found. This results in an apparent slight improvement over the recall values shows in Table 4 (which are repeated here for convenience), although none of the apparent improvements turned out to be statistically significant. But the key point for e-discovery practice is that the potential gains are small, and of course costs would grow with the number of stages in the query negotiation process. Our results point to a strong benefit to one stage of renegotiation, and at present we can find no basis for suggesting that a second renegotiation would typically be worthwhile.

5. CONCLUSIONS AND FUTURE WORK

E-discovery is one example of a class of search problems in which disclosure must be balanced with confidentiality. Professional practice in the era of paper was for the most part

Target	Limit	1K	2K	5K	10K	25K
25	2,000	0.7225	0.8199	0.9481	0.9711	0.9842
25	500	0.7871	0.8784	0.9351	0.9596	0.9797
1	N/A	0.8030	0.8920	0.9500	0.9740	0.9862

Table 5: Comparing alternative cutoffs for the maximum number of documents to review.

a manual search-then-review process, and this same process has now been widely adopted for digital content as well with only limited use of the capabilities of modern search engines. This kind of direct transfer is natural early in the life of any new technology, but the next natural stage in the process of innovation is to start from there to explore alternatives that might be better suited to new situations. Our investigation in this paper of selective disclosure for known-topic searching is one such path that an exploration might take. We have found that one round of query renegotiation can improve retrieval effectiveness (for a fixed level of review effort) and that it would likely be helpful to have a substantial number (in our experiments, 25 or above) of relevant documents before renegotiating the query. That gains are possible is, of course not at all surprising—simulations of interactive relevance feedback have long been known to show improvements over initial queries (at least on average across many topics) [5]. Our work's focus on deep recall rather than early precision, motivated by the primacy of recall in many e-discovery settings, add a useful perspective to that body of work. Moreover, somewhat surprisingly, our experiments offer no evidence that additional rounds of selective disclosure and query renegotiation are helpful. This might reflect our limited ability to model query renegotiation using fully automated techniques, or it may be an artifact of the limited number of known relevant documents for the topics in our collection (131, on average), or it may well be a real effect that would be observed in real e-discovery settings.

That suggests two very obvious next steps: (1) repeat these experiments using additional test collections, and (2) test any results that are confirmed in other settings with user studies. The TREC legal track test collection would be a natural focus for one replication of this study, perhaps as part of the 2008 legal track relevance feedback task. The more complex queries in that collection (with Boolean, proximity and truncation operators), and the complexity added by the variable OCR quality and the nature of that collection's metadata, pose challenges that will need to be addressed, however. It might, therefore, also be useful to first replicate our study design more directly using a test collection with a simpler structure (for which TREC and TREC-like evaluations around the world offer numerous possibilities).

Designing an affordable user study is an even larger challenge, of course, but the key to affordability will be to focus the research questions narrowly. Our results should help to inform that study design. For example, if upon replication it again turns out that only a single round of selective disclosure is needed, then a first user study could test whether the gains we see from one round can indeed be achieved with actual negotiations between representative users. The human subject costs in this case would be limited to the renegotiation process – subsequent scoring could be fully automatic. If those results are promising, a second excursion to explore whether human subjects can obtain benefits from a second round of selective disclosure could be conducted, again with automatic scoring. And once a promising protocol was in this way, that protocol could be subjected to a more complete vetting using new relevance judgments, perhaps as part of the TREC legal track's interactive task.

Of course, the ultimate extent to which our model of selective disclosure can influence practice in the real world depends on a number of factors that extend well beyond mere technical issues. In particular, adoption of such a process largely seems dependent on the degree to which the legal profession is willing to embrace notions of collaboration and transparency on such matters.

Acknowledgments

The authors are grateful to Bruce Hedin, Dave Lewis, Paul Thompson, and Stephen Tomlinson for helping to organize the TREC legal track that inspired this work.

6. REFERENCES

[1] J. R. Baron, D. D. Lewis, and D. W. Oard. TREC-2006 legal track overview. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*, 2007.

[2] D. C. Blair and M. E. Maron. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun. ACM*, 28(3):289–299, 1985.

[3] D. W. Oard, J. R. Baron, S. Tomlinson, and J. R. Baron. TREC-2008 legal track overview. In *Proceedings of the Seventeenth Text REtrieval Conference (TREC 2006)*, 2008.

[4] G. L. Paul and J. R. Baron. Information inflation: Can the legal system adapt? *Richmond Journal of Law & Technology*, 13(3), 2007.

[5] K. Sparck Jones. Search term relevance weighting given little relevance information. *Journal of Documentation*, 35(1):30–48, 1979.

[6] K. Sparck-Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments part 2. *Information Processing & Management*, 36(6):809–840, 2000.

[7] S. Tomlinson, D. W. Oard, J. R. Baron, and P. Thompson. Overview of the TREC 2007 legal track. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*, 2008.

[8] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, 2000.

[9] E. M. Voorhees. Overview of the TREC 2005 robust retrieval track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, 2006.

[10] F. C. Zhao, Y. Lee, and D. Medhi. Evaluation of query formulations in the negotiated query refinement process of legal e-discovery: UMKC at TREC 2007 legal track. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*, 2008.