The Search Problem Posed By Large Heterogeneous Data Sets In Litigation: Possible Future Approaches To Research

Jason R. Baron
Director of Litigation
Office of the General Counsel
National Archives and Records Administration
8601 Adelphi Road, Suite 3110
College Park, MD 20740
Tel. 301-837-1499
Jason.baron@nara.gov

Paul Thompson
Research Associate Professor
Department of Computer Science
Dartmouth College
6211 Sudikoff Laboratory
Hanover, NH 03755
Tel. 603-646-8747
Paul.Thompson@dartmouth.edu

ABSTRACT

Lawyers and their large institutional clients increasingly face the enormous problem of how to efficiently and efficaciously conduct searches for relevant documents in large heterogeneous electronic data sets, for the purpose of responding to litigation demands. Past research indicates that lawyers greatly overestimate their true rate of recall in civil discovery. The unprecedented size, scale, and complexity of electronically stored data now potentially subject to routine capture in litigation, for purpose of preservation, access, and review, presents information retrieval researchers with a series of important challenges to overcome. This paper describes the current context of e-discovery and discusses the potential for IR and AI research to address the challenges of conducting e-discovery. The TREC Legal Track is presented as a forum for the evaluation of e-discovery research and one new evaluation measure, elusion, is described, which has potential for addressing problems of measuring recall.

Categories and Subject Descriptors

H.3.3 Information Search and Retrieval

General Terms

Experimentation

Keywords

E-Discovery, Evaluation.

INTRODUCTION

As stated elsewhere at ICAIL 2007, lawyers and their large institutional clients increasingly face the enormous problem of how to efficiently and efficaciously conduct searches for relevant documents in large heterogeneous electronic data sets, for the purpose of responding to litigation demands [2]. The potential magnitude of the search problem is highlighted by past research [4] 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"). The unprecedented size, scale, and complexity of electronically stored data now potentially subject to routine capture in litigation, for purpose of preservation, access, and review, 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 corpora of documents aimed at evaluating the efficacy of the search task [4, 23], each ended up involving a pooling size of approximately 3×10^4 documents, constituting the size of the collection which was made subject to a human assessment process determining "responsiveness" on the item level. (In TREC, this pool was drawn from an overall universe of 6×10^6 documents.) For good reasons, these past efforts have utilized certain design and/or evaluation criteria that may or may not prove to be optimal for future research projects involving data sets with much higher orders of magnitude of both responsive and non-responsive documents. But as data sets get larger, "indeterminacy multiplies making it increasingly difficult to conduct successful specific or exhaustive searches." [5]. 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?

As an initial step in thinking about how to structure IR research for the purpose of advancing our knowledge on improving the

efficacy of legal searches in a real world context, three types of relevant factors potentially serve to inform the discussion, as set out in Part A, *infra*: (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. In Part B we review IR and AI research as it has been historically applied to the legal domain and propose how it can now be applied to e-discovery. With these initial considerations in mind, in Part C we briefly describe the parameters, and some preliminary results, of what has been Year 1 of the current TREC Legal Track, as a baseline for future research, before proceeding with a discussion in Part D of evaluation challenges for future e-discovery research.

A. THE LEGAL CONTEXT

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 electronically stored information ("ESI"). Directly as the result of the unparalleled volume and nature of such newly arising forms of evidence, Congress and the Supreme Court approved recent changes to the Federal Rules of Civil Procedure, in effect as of December 1, 2006, which *inter alia* add "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*,[1] the recently decided 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 an automated search methods with rudimentary Boolean search terms, the government uncovered 200,000 emails (with attachments), in need of 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 [3]. 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 [14]. 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 relevancy. Thus, greater reliance on all forms of automated methods will of necessity be commonplace throughout the profession, in turn raising questions of their accuracy, efficacy, and completeness. as a measure of recall, precision, or against any agreed upon statistical metric.

In addition to exponential increases in volume, the data sets themselves are rapidly evolving. The past decade has seen not only explosion in email traffic, but also the growth of dynamic databases of all kinds, including universes of data found on the Web and on corporate intra-nets, including wiki collaborations and the blogosphere, as well as instant messaging and various forms of audio and video formats. 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.

Today's full "corporate desktop" menu of heterogeneous data constitutes a prime target in civil litigation, and thus is a worthy candidate of further academic research evaluating the efficacy of search methods applied both to textual documents as well as information stored in other forms of rich media.

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 "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.

How Legal Searches Are Actually Performed. The state of practice as it exists today consists of lawyers, in response to broadly worded discovery demands, directing 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. Courts have supported the use of keywords as a fair way in which to sample if not fully respond to discovery demands, in some cases insisting that the parties cooperate through negotiations over what keywords to input in large databases.[2] No case law is known to exist, however, in which parties adjudicated the use of any of the various well-known alternative forms of search methods extant, *e.g.*, utilizing algebraic or probabilistic means of searching as an alternative to set-based Boolean inquiries. The legal field in this regard is 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 "all" responsive documents in vast data sets.

In most instances, lawyers place greatest emphasis on finding responsive documents, thus the emphasis on measures of recall

[1] 2006 WL 2380648 (D.D.C. 2006).

[2] See, e.g., *Zubulake v. U.B.S. Warburg*, 229 F.R.D. 422 (S.D.N.Y. 2004); *Treppel v. Biovail Corp.*, 233 F.R.D. 363 (S.D.N.Y. 2006); and *Balboa Threadworks v. Stucky*, 2006 WL 763668 (D. Kan. 2006). See generally Sedona Conference [21].

over precision. As a secondary measure, lawyers would find valuable search methods that maximize efficiency in eliminating false positives found through automated search methods – but not generally at the expense of finding a lesser amount of relevant documents. Thus, one could expect the following type of calculus in a legal context:

Assume a document collection of 30 million documents. Further assume that using Automated Search Method A, one encounters 200,000 "hits," i.e., possibly relevant documents, which in turn would take six months of manual review by a team of lawyers and assistants to weed out nonresponsive false positives, leaving a total of 100,000 relevant documents.

Against the same test collection, assume Search Method B yields 100,000 "hits," which in turn would take only three months to review by the same manual team of individuals, but where this further review process results in a total of 90,000 relevant documents.

In this extreme case, while Search Method B is vastly more efficient measured as a matter of precision, arguably most lawyers in civil litigation – all other factors being equal -- would fail to choose Method B over Method A, out of concern that due diligence requires a review of the "missing" 10,000 relevant documents found by Method A. Hence, the primary focus on measures of recall.

The recall/precision trade-off has neither been discussed in case law nor is it well-articulated or even understood by the legal community.

B. APPLYING IR AND AI RESEARCH TO THE NEW E-DISCOVERY PROBLEM

Going back to the 1890s with the development of Shepardizing, the legal community has been at the forefront of advanced search technology. When full text retrieval first became commercially available in the 1970s, again the legal community pioneered its development. Lexis-Nexis and Westlaw supported full text search and the full text search STAIRS product from IBM was used in the large litigation studied by Blair and Maron [4]. Similarly, the AI and Law community has for decades applied advances in AI research to the legal domain.

While ranked retrieval had been the focus of academic research on information retrieval since the 1960s [15, 20], it was not until the early 1990s that West Publishing's Westlaw legal search engine introduced ranked retrieval to large scale commercial online retrieval by offering a ranked retrieval search mode. A year later Lexis-Nexis also provided a ranked retrieval mode for its users, as did Dialog, another large proprietary system, though one not specifically serving the legal community,

Simultaneous with these developments in the legal domain, in the early 1990s the U. S. government funded the Tipster program [10] to support the research and development of revolutionary new algorithms for both information retrieval and natural language understanding, specifically information extraction. One of the Tipster contractors was the University of Massachusetts, Amherst, with its probabilistic retrieval system, Inquery [24]. A version of the Inquery system, tailored for the legal domain, became the ranked retrieval mode of the Westlaw system.

What happened next? The legal community never adopted ranked retrieval searching. The overwhelming majority of users of Westlaw and Lexis-Nexis continued to use the traditional Boolean search mode, instead of the new ranked retrieval mode. Meanwhile, search technology outside the legal community continued to evolve. The Tipster program continued through most of the 1990s and then was succeeded by related government programs such as Translingual Information Detection Extraction and Summarization (TIDES). The Message Understanding Conference [12] which corresponded to the information extraction component of Tipster, was also succeeded by the Automatic Content Extraction (ACE) program [1]. New related programs also emerged such as Novel Intelligence in Massive Data (NIMD) and Advanced Question Answering for Intelligence (AQUAINT), both sponsored by the Advanced Research and Development Activity (ARDA).

Commercial retrieval in the legal domain, whether that of services such as Westlaw or Lexis-Nexis, or of domain-specific vendors, e.g., for litigation support, continued to provide similar services to those that they had been providing. Many of the new technologies which have been developed over the past few years by intelligence community programs such as TIDES, NIMD, and AQUAINT, and most recently the DARPA Global Autonomous Language Exploitation (GALE) program [11] can be applied to the legal domain, especially to e-discovery. The data sets of interest to the intelligence community have many of the same characteristics as those of interest in e-discovery. The Enron e-mail data set provides a good example, both in terms of its type of content being of interest to both the intelligence and e-discovery communities and in terms of the different ways in which it would be, or has been, used as an evaluation tool by each community. This data set was considered for use with the TREC Legal Track. At the same time it has been used as a surrogate for a terrorism database. The Society for Industrial and Applied Mathematics (SIAM) 2005 International Conference on Data Mining Workshop on Link Analysis Counterterrorism, and Security encouraged participants to use the Enron data set in this way [22]. Also in 2005 the International Conference on Artificial Intelligence and the Law (ICAIL) sponsored a similar workshop on Data Mining, Information Extraction, and Evidentiary Reasoning for Law Enforcement and Counter-Terrorism [13]. For the SIAM workshop each participant was encouraged to present research using the Enron collection. About half of the participants did. Those who did each analyzed the data differently. Had the Enron collection been used for evaluation for the TREC Legal Track, it would have been necessary, just as was done with the CDIP collection (see Part C), to develop topics or queries, and to provide relevance judgments. Both of these tasks require specialized expertise.

A variety of other conferences have emerged that have supported various aspects of information retrieval, information extraction, or text mining to various types of data related to e-discovery, e.g., the Conference on E-Mail and Anti-Spam [8]. Research on authorship attribution and sentiment detection has attracted

increased attention. GALE has provided new research funding for natural language processing.

The AI and Law community has continued to address information retrieval, information extraction and related research areas such as textual case-based reasoning and question answering, e.g., Weber et al. [27] Moens [16], and Branting [6]. Much of this research has focused on legal texts such as case law documents and statutes. E-discovery opens a wider domain of heterogeneous data and document types to which advanced technology developed in the AI and Law Community can be applied. While the TREC Legal Track has focused to date on textual collections, the scope of e-discovery also includes multimedia data types of the sort studied in the government programs mentioned above, e.g., GALE.

During this time, neither the world nor the legal tech sector has stood still. As e-discovery issues have come in to prominence with the development of a substantial body of case law (even before the 2006 rules changes), a legal tech 'colossus' has grown in parallel. According to one leading marketplace study, corporate America is expected to spend $ 1.4 billion dollars in 2006 on e-discovery. with the figure growing to $ 4.8 billion by 2011 [17]. Given the continued exponential increases in volume of ESI, all of these factors should contribute to a sunny forecast for actors in the legal marketplace who purport to be selling a better mousetrap, i.e., making a claim to have better search and retrieval methods, systems and tools for more robust and efficient searches.

Interestingly, however, one can search in vain through a vast amount of proprietary literature without citation or grounding to AI or IR research; nor do tech firms typically proffer in-house studies of a proprietary nature showing the efficacy of their products. This was most recently confirmed by the authors' joint attendance at Legal Tech 2007, a very well-attended event held each year in Manhattan in January. During the trade show, we attempted on an informal basis, but without success, to obtain benchmarking studies from various leading legal tech company representatives, showing some form of testing or evaluation of their products against objective criteria. The authors' anecdotal experience only serves to confirm that a very important and substantial research void exists waiting to be filled.

C. THE TREC 2006 LEGAL TRACK

TREC's purpose is to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. In particular, the TREC workshop series has a stated goal of encouraging research in information retrieval based on large test collections. [26]. As part of the Fifteenth TREC year sponsored by the National Institute of Standards and Technology (NIST), a new "legal track" was introduced, consisting of what has now become a multi-year collaborative information retrieval research project focused on e-discovery applications. A key goal of the legal track has been to attempt to use objective benchmark criteria in comparing search technologies as used in a setting modeled on how lawyers actually undertake real discovery.

Utilized for the TREC legal track was the Complex Document Information Processing (CDIP) Test Collection from the University of Illinois, which includes 6.9 million documents released by tobacco companies in connection with a "Master Settlement Agreement" with several state attorneys general. For use against this collection, five hypothetical complaints were created, accompanied by 43 topics in the form of "requests to produce." For each topic, a lawyer representing the responding party initiated a proposed Boolean set, which after further back and forth negotiations became the baseline "negotiated Boolean query" for use by participants submitted runs. Six participating institutions submitted a combined total of 31 "fully automatic" runs, where runs were requested to depth 5,000 (i.e., up to 5,000 documents deemed relevant by any particular system). A manual searcher separately engaged in iterative searches against the test collection to come up with up to 100 documents deemed by her to be relevant per topic, to add to the NIST pool. NIST thereafter created judgment pools based on the results of the combined runs. Thereafter, 35 volunteer assessors manually judged the relevance of approximately 800 documents contained within each topic pool from the automatic runs. Based on the first year assessment undertaken, 32,738 topical relevance judgments are now available for 40 use cases." [23]. Some additional dual assessment was undertaken.

An initial condition placed on the TREC legal track design was that topics be chosen with low "R" (responsive) values, to the extent knowable, *i.e.*, that experimental surveys of the topic collection conducted by track designers would result in artificial limitation of the true number of "responsive" documents to a set amount, generally between 1 and 1000 hits for the baseline Boolean negotiated search. As it turned out, the "Boolean constraint," i.e., the number of documents ("B") found by the baseline Boolean run using the negotiated queries, ended up varying across the range of topics from B=1 to B=128,195 (where 23 topics had B less than or equal to 5,000). Importantly, for purposes of result reporting, the TREC research program assumes that there will be no responsive documents in the larger data universe represented by the topic collection, here CDIP. As noted, however, "[t]he use of pooling to produce a test collection has been questioned because unjudged documents are assumed to be not relevant. Critics argue that evaluation scores for methods that did not contribute to the pools will be deflated relative to methods that did contribute because the non-contributors will have highly ranked unjudged documents." [25], at p. 4.

Research results from Year 1 of the TREC Legal Track have been set out at some length in a TREC 2006 Legal Track Overview paper [23], to which interested readers are respectfully referred. Two of the main results of the research will be synopsized here. First, as shown below in Figure 1, approximately 57% of the known relevant documents across all topics were found to by the baseline Boolean query (either uniquely by the Boolean query or by the reference Boolean and also one or more other systems).

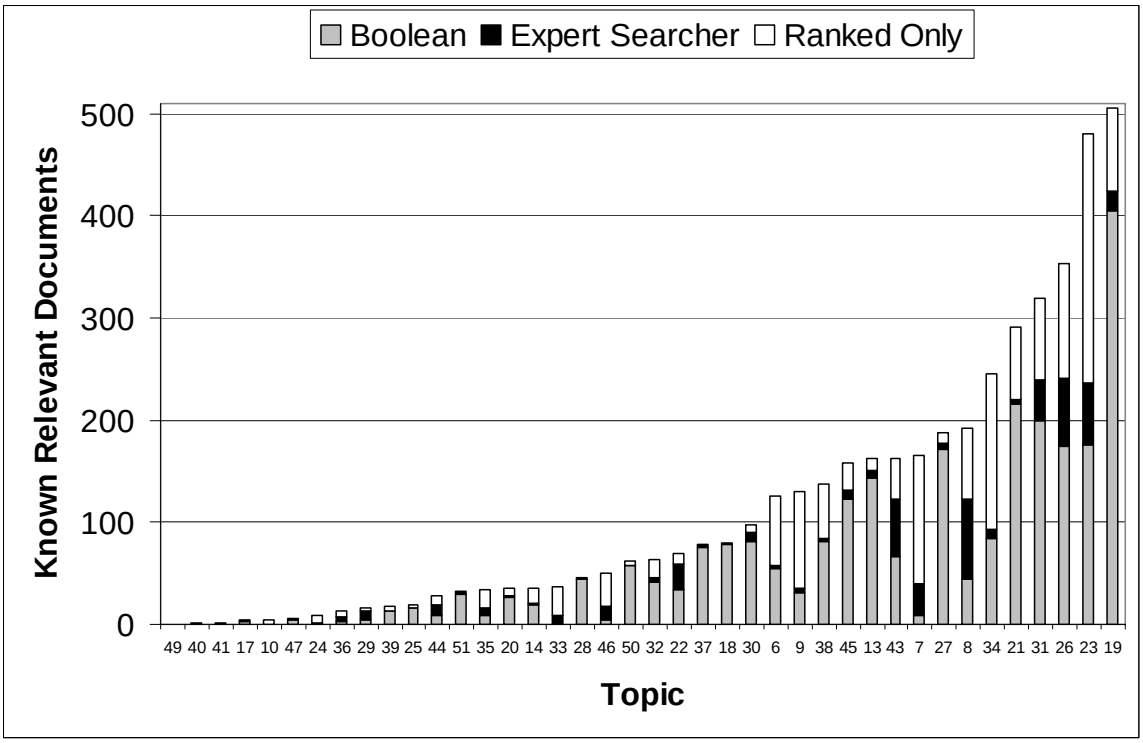

Figure 1. Known relevant documents found by the Reference Boolean system (grey), found by the expert searcher but not the reference Boolean system (black), and found uniquely by at least one other system (white).

As a corollary, a substantial proportion (coming to 32%) of the overall number of known relevant documents were found only by ranking systems *other* than the Boolean runs or the manual searcher's efforts. (The manual searcher found the remaining 11% of the overall number of unique relevant documents.)

In turn, Figure 2 constitutes a separate comparison of Boolean and ranked runs using an R-precision measure yielded inconclusive results. Figure 2 compares the ranked retrieval runs using R-precision, a precision-oriented measure that is widely used, understood, and reported. R-precision is computed as the average across topics of the relative frequency of relevant documents in the top R ranks (where R is the number of known relevant documents for that topic). The five bars show the best scoring runs from a manual searcher, a reference Boolean run, and the top three participating teams. Because R-precision is focused early in the ranked list, this measure would be expected to favor ranked retrieval systems. All four Boolean runs were, however, ranked in some way after being subjected to the Boolean constraint. The result is, therefore, in some sense fair in those cases.

Three results are clearly evident in this data. First, the best runs from the three participating systems shown here were nearly indistinguishable by the R-precision measure, and one of those three was subjected to a Boolean constraint. Indeed, the reference Boolean run did about as well in this high-precision region as the best unconstrained ranked retrieval runs. This is notable because Boolean runs can retrieve only documents that satisfy the Boolean query, while the ranked runs had no such constraint. A second result is that Boolean systems are not all created equal—two of the four Boolean runs did about twice as well (by this precision-oriented measure) as the other two. In one case this appears to result from using the initial rather than the final negotiated Boolean queries. In the other case the differences appear to result from incomplete support for extended Boolean operators. Third, the expert manual searcher's results were (by this measure) noticeably better than any fully automatic system. This is particularly notable because ranking the result set was not a part of the expert manual searcher's task, and thus further improvements may have been possible. This suggests that focusing some further attention on interactive evaluation might yield interesting results.

As per above, all of the results from Year 1 of the TREC Legal Track have been posted on the NIST TREC web site prior to ICAIL 2007. The evaluation measures undertaken in the TREC legal track are expected to make an important contribution to better understanding of Boolean methods vs. nonBoolean ranking schemes across a variety of legal settings. However, further research efforts and modeling of real world-litigation beyond TREC is arguably in order if litigation involving the recovery of millions of potentially responsive documents is a legitimate and growing concern, as it would appear to be.

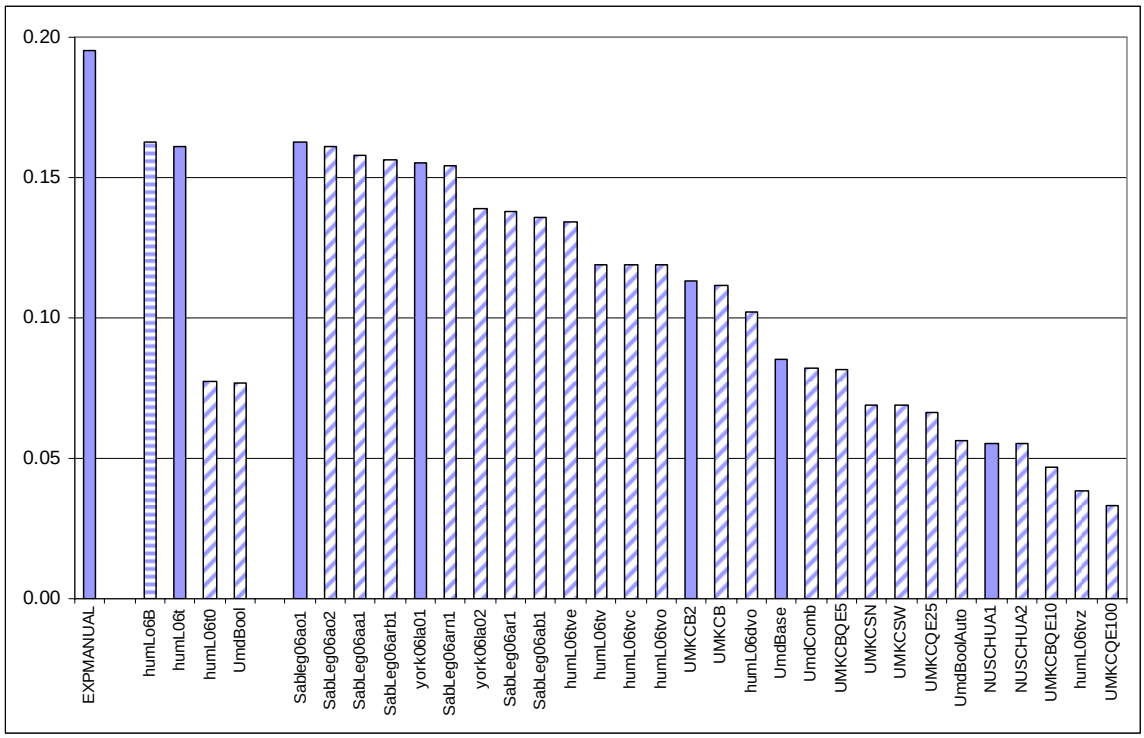

Figure 2. Mean precision at R (the actual number of known relevant documents for each topic). Ranked runs on left side, Reference runs on right side. Best run for each team shown as solid bar.

D. Evaluation Challenges for Future Research Related to e-Discovery Posed by Large Heterogeneous Data Sets In Litigation

We have described the search problems posed in large heterogeneous data sets in litigation and the initial attempt to address these problems by the 2006 TREC Legal Track. Apart from finding new and better ways in which to search large collections there is also a need to develop a new evaluation methodology for such large collections. The pooled relevance assessment methodology, which assumes that a pool including the top *n* ranked results of each participating retrieval system provides a representative sample of the test collection as a whole, has been shown to lead to biased results with larger collections now being evaluated [7].

The experience with the 2006 TREC Legal Track has suggested that one major challenge facing the design of future e-discovery evaluations is the need for better evaluation measures for high-recall effectiveness. When TREC was first begun in the early 1990s, researchers were concerned with the limitations of existing information retrieval text collections. In particular the small size of these collections was a concern. TREC addressed these concerns by developing much larger test collections. However, with these larger collections it was no longer possible, as had been done in the past with some test collections, to review each query document pair in the collection for relevance. The assumption was made that if a large number of systems participated in TREC that by judging for relevance the top *n* documents retrieved by each system this pooled set of judged documents would provide a representative sample of the document collection as a whole, particularly if many systems with diverse ranking algorithms participated. This assumption seemed reasonable, but is now breaking down as collections are becoming increasingly large. The problem is made worse if there are also a large number of relevant documents for each query. For example, assume that a document collection has one billion documents and that around one million of these may be relevant for a given query. Judging the top several hundred documents retrieved for each query for each of 20 or 30 participating systems will not give a good estimate of recall for the collection as a whole.

One further measure that has been advanced has been termed “elusion” [19]. As described by Roitblat, rather than estimating the proportion of responsive documents that have been retrieved, one determines whether significant numbers of documents were missed by the retrieval process. To assess elusion, one evaluates a randomly selected set of nonretrieved documents. If there are any responsive documents in this sample one adjusts the retrieval criteria so that these documents would be retrieved. Then a new random sample is drawn.

Elusion can be estimated from a random sample of nonretrieved documents. The larger the sample, the more accurate the estimate. The optimal sample size depends on the confidence level desired and on the desired maximum probability of

nonresponsive documents among the nonretrieved set. In this example a confidence level of 0.98 is chosen. The number of documents that must be sampled is determined by the formula provided in [19]:

$$n = \frac{\log(\alpha)}{\log(1 - p_s)} = \frac{\log(0.02)}{\log(1 - 0.02)} \approx 200$$

In this example, the maximum prevalence of responsive documents in the nonretrieved set is set to be 2%. No more than 2% of the rejected documents are expected to be responsive to (p_s). Based on these assumptions, 200 documents are randomly selected from those that were not retrieved. If any of those documents are found to be responsive, the search criteria are revised to capture those responsive documents and a new sample of 200 documents is selected. This process is repeated until the sample comes up with 0 responsive documents. Rather than merely estimating our level of success, as would recall, elusion allows us to assess whether our entire process has succeeded to the required level.

The success of the elusion measure is dependent on starting assumptions regarding the prevalence of responsive documents in a large heterogeneous universe. Over time, lawyers in real-world contexts would be expected to become more familiar with rates of responsiveness in large data universes, after having used automated means of search as "first approximations" before proceeding with manual review. It would therefore be extremely useful to test elusion measures in a variety of real-world settings, where the overall rate of responsiveness may vary widely between significantly under 1% of a collection as a whole, to many multiples of 1 or 2%, for the purpose of establishing the method as a satisfactory evaluation measure.

More generally, lawyers are in need of greater assistance from the AI and IR communities in coming up with ways in which the "reasonableness" of Boolean and non-Boolean search methods are to be evaluated. What passes today for anecdotal evidence and unfounded assumptions on the completeness of status quo search methods must yield to the development of metrics that present judges with objective ways in which to interpret reasonableness of search methods used in particular litigation settings.

E. CONCLUSION

In light of looming increases in the volume of information to be processed in civil litigation, the legal community is in need of new and better methods and technologies aimed at maximizing the efficiency of the search process used in discovery. The absence of objective benchmarking of comparative approaches is a problem that is partially addressed by ongoing research in the TREC legal track. One potential evaluation measure, elusion, aims to better capture whether responsive documents have been missed by the choice of search method used. However, with increased awareness of the special needs of lawyers in connection with e-discovery, a wide variety of other approaches drawn from AI efforts in other contexts may yet serve as suitable candidates for future research directly tied to the types of heterogeneous data and document types typically arising in real-world litigation contexts.

REFERENCES

[1] ACE, http://www.nist.gov/speech/tests/ace/

[2] Baron, J. R., with Oard, D.. M., Lewis, D. L., and Thompson, P. Supporting Search and Sensemaking in Electronically Stored Information for Discovery *Proceedings*, ICAIL 2007 Workshop Description, http://www.umiacs.umd.edu/~oard/desi-ws/.

[3] Baron, J. R., Toward A Federal Benchmarking Standard for Evaluating Information Retrieval Products Used in E-Discovery, 6 *Sedona Conference Journal* (2005), 237-246 (available on Westlaw, Lexis)

[4] Blair, D. C., and Maron, M. E., An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System, *Communications of the ACM* 28, 3 (1985), 289-299.

[5] Blair, D., *Wittgenstein, Language and Information: 'Back to the Rough Ground!,'* Springer, Dordrecht, The Netherlands, 2006.

[6] Branting, K., The Role of Syntactic Analysis in Textual Case Retrieval, *ICCBR Workshops* 2005.

[7] Buckley, C., Dimmick, D., Soboroff, I., and Voorhees, E. Bias and the Limits of Pooling. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '06) (Seattle, August 6-11, 2006).ACM Press, New York, NY, 2006, 619-620.

[8] CEAS. 2007, http://www.ceas.cc/

[9] Collaborative Expedition Workshop #45, *Advancing Information Sharing, Access, Discovery and Assimilation of Diverse Digital Collections Governed by Heterogeneous Sensitivities*, held Nov. 8, 2005, see http://colab.cim3.net/cgi-bin/wiki.pl?AdvancingInformationSharing_DiverseDigitalCollections_HeterogeneousSensitivities_11_08_05

[10] DARPA *Proceedings Tipster Text Program Phase III October 1996 - October 1998*.

[11] GALE, 2007. http://www.darpa.mil/ipto/Programs/gale/index.htm

[12] Grishman, R. and Sundheim, B., Message Understanding Conference - 6: A Brief History *COLING 1996*, 466-471.

[13] ICAIL 2005. http://dale.liacs.nl/resources/misc/icail2005.pdf

[14] Jensen, J.H., Special Issues Involving Electronic Discovery, 9 *Kansas Journal of Law and Public Policy* 425 (2000) (available on Westlaw, Lexis).

[15] Maron, M. E. and Kuhns, J. L. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7, 3, (1960), 216-244.

[16] Moens, M.-F., Combining Structured and Unstructured Information in a Retrieval Model for Accessing Legislation *ICAIL 2005*.

[17] Murphy, B., Special Issue: Readings on Case-based reasoning: Believe It — eDiscovery Technology Spending To Top $4.8 Billion By 2011, http://www.forrester.com/Research/Document/Excerpt/0,7211,40619,00.html.

[18] Paul, G. L., and Baron, J. R. Information Inflation: Can The Legal System Adapt?, 13 *Richmond Journal of Law and Technology* 10 (2007), see law.richmond.edu/jolt/v13i3/article10.pdf

[19] Roitblat, H., Appendix A to The Sedona Conference Draft Commentary on Search and Retrieval, Version 1.0 (April 2006) (unpublished draft available from the authors).

[20] Salton, G. Manipulation of Trees in Information Retrieval. *Communications of the ACM*. 5, 2 (1962), 103-114.

[21] The Sedona Conference, *The Sedona Principles: Best Practices Recommendations & Principles for Addressing Electronic Document Production* (2005 version), see http://www.thesedonaconference.org/content/miscFiles/publications_html

[22] SIAM, 2005. *International Conference on Data Mining Workshop on Link Analysis, Counterterrorism and Security*. http://www.cs.queensu.ca/~skill/siamworkshop.html

[23] TREC 2006 Legal Track Overview, Baron, J. R., Oard, D. W., Lewis, D. D. *available at* http://trec-legal.umiacs.umd.edu/

[24] Turtle, H. and Croft, W. B. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9, 3, (1991), 187-222.

[25] Voorhees, E., Overview of TREC 2005, *The Fourteenth Annual Text Retrieval Conference* (NIST 2005)

[26] Voorhees, E. and Harmon, D. K. (eds.), *TREC: Experiment and Evaluation in Information Retrieval* (Digital Libraries and Electronic Publishing: 2005).

[27] Weber, R. O., Ashley, K., Bruninghaus, Textual Case-based reasoning**.** *Knowledge Engineering Review, Special Issue: Readings on Case-based reasoning* (2006).