Manuscript accepted for publication in the Richmond Journal of Law and Technology, Volume XVII, Issue 3 (Spring 2011).

Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review

Maura R. Grossman, J.D., Ph.D.
Wachtell, Lipton, Rosen & Katz[1]

Gordon V. Cormack, Ph.D.
University of Waterloo

Abstract

E-discovery processes that use automated tools to prioritize and select documents for review are typically regarded as potential cost-savers – but inferior alternatives – to exhaustive manual review, in which a cadre of reviewers assesses every document for responsiveness to a production request, and for privilege. We offer evidence that such technology-assisted processes, while indeed more efficient, can also yield results superior to those of exhaustive manual review, as measured by recall and precision, as well as F_1, a summary measure combining both recall and precision. Our evidence derives from an analysis of data collected from the TREC 2009 Legal Track Interactive Task. We show that, at TREC 2009, technology-assisted review processes enabled two participating teams to achieve results superior to those that could have been achieved through a manual review of the entire document collection by the official TREC assessors.

[1] The views expressed herein are solely those of the author and should not be attributed to her firm or its clients.

Maura R. Grossman is Counsel at Wachtell, Lipton, Rosen & Katz, where she advises the firm and its clients on legal, technical, and strategic issues involving electronic discovery and information management, both in the U.S. and abroad. She has represented Fortune 100 companies and major financial services institutions in corporate and securities litigation, including both civil actions and white-collar criminal and regulatory investigations. Ms. Grossman has been appointed by the Chief Administrative Judge to serve as co-chair of the E-Discovery Working Group advising the New York State Unified Court System, and is involved in other initiatives to provide education on e-discovery to federal and state court judges. Since 2010, Ms. Grossman has served as a coordinator of the Legal Track of the National Institute of Standards and Technology's Text Retrieval Conference ("TREC"), a joint government/industry/academic research project studying the application of automated information retrieval technologies to e-discovery. She is as an adjunct professor at the Rutgers School of Law – Newark and Pace Law School, where she teaches courses on e-discovery. Ms. Grossman also is a member of The Sedona Conference® Working Groups on Electronic Document Retention and Production, and on International Electronic Information Management, Discovery and Disclosure. She assisted in editing *The Sedona Conference® Commentary on Achieving Quality in E-Discovery* (May 2009), and serves on the Advisory Boards of BNA's *Digital Discovery and E-Evidence Report* and the Georgetown University Law Center's Advanced E-Discovery Institute. In addition to her law degree from the Georgetown University Law Center, Ms. Grossman holds an M.A. and Ph.D. in Clinical/School Psychology from Adelphi University.

Gordon V. Cormack holds the rank of Professor in the David R. Cheriton School of Computer Science at the University of Waterloo in Canada. Prof. Cormack is co-director of the Information Retrieval Group at Waterloo, and co-author of *Information Retrieval: Implementing and Evaluating Search Engines* (MIT Press 2010). Prof. Cormack is the author or co-author of more than 100 peer-reviewed articles on information retrieval, computer systems, and computer education. Prof. Cormack's current research interests involve the application of machine learning technologies to critical applications like spam filtering and e-discovery, as well as the scientific evaluation of the effectiveness of such methods. For more than 12 years, Prof. Cormack has been a program committee member of TREC at large. From 2005 through 2007, he was coordinator of the TREC Spam Track, and since 2010, he has been a coordinator of the TREC Legal Track. Prof. Cormack is a winner of the INFORMS 2009 Data Mining Contest and the ECML/PKDD 2006 Discovery Challenge. He was the Scientific Director of the 2010 International Olympiad in Informatics, an international computer science contest for high school students from 73 countries, which was featured in the December 2010 issue of *Wired*. Prof. Cormack has coached Waterloo's ACM International Collegiate Programming Contest team for the past 14 years, qualifying for the World Championships each year, and winning the World Championship once, and the North American Championship three times.

1 Introduction

The Sedona Conference® Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery [2] cautions that:

> [T]here appears to be a myth that manual review by humans of large amounts of information is as accurate and complete as possible – perhaps even perfect – and constitutes the gold standard by which all searches should be measured. Even assuming that the profession had the time and resources to continue to conduct manual review of massive sets of electronic data sets (which it does not), the relative efficacy of that approach versus utilizing newly developed automated methods of review remains very much open to debate.

While the word *myth* suggests disbelief, the literature contains little scientific evidence to support or refute the notion that automated methods, while improving on the efficiency of manual review, yield inferior results. This work presents evidence supporting the contrary position: that a technology-assisted process, in which only a small fraction of the document collection is ever examined by humans, can yield higher recall and/or precision than an exhaustive manual review process, in which the entire document collection is examined and coded by humans.

A *technology-assisted review process* involves the interplay of humans and computers to identify the documents in a collection that are responsive to a production request, or to identify those documents that should be withheld on the basis of privilege. A human examines and codes only those documents that are identified by the computer – a tiny fraction of the entire collection. Using the results of this human review, the computer codes the remaining documents in the collection for responsiveness (or privilege). A technology-assisted review process may involve, in whole or in part, the use of one or more approaches including, but not limited to, keyword search, Boolean search, conceptual search, clustering, machine learning, relevance ranking, and sampling.[3] In contrast, *exhaustive man-*

[2] The Sedona Conference® Working Group on Best Practices for Document Retention and Production (WG1), *The Sedona Conference® Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 Sedona Conf. J. 189, 199 (2007), *available at*
http://www.thesedonaconference.org/content/miscFiles/Best_Practices_Retrieval_Methods___-revised_cover_and_preface.pdf.

[3] *See, e.g.*, Sedona, *supra* note 2, at 217*;* Stefan Büttcher, Charles L. A. Clarke & Gordon V. Cormack, *Information Retrieval: Implementing and Evaluating Search Engines* (2010). The

ual review requires one or more humans to examine each and every document in the collection, and to code them as responsive (or privileged) or not.

A review of the literature (*infra* Section 2) suggests that manual review is far from perfect. Recent results from The Text Retrieval Conference ("TREC"), sponsored by the National Institute of Standards and Technology ("NIST"), show that technology-assisted processes can achieve high levels of recall and precision.[4] By analyzing data collected during the course of the TREC 2009 Legal Track Interactive Task,[5] we demonstrate that the levels of performance achieved by two technology-assisted processes exceed those that would have been achieved by the official TREC assessors – law students and lawyers employed by professional review companies – had they conducted a manual review of the entire document collection.

2 Context

Under Rule 26(g)(1) of the Federal Rules of Civil Procedure, an attorney of record must certify, "that to the best of the person's knowledge, information, and belief formed after a reasonable inquiry," that every discovery request, response, or objection is "consistent with these rules . . . [,] not interposed for any improper purpose, such as to . . . cause unnecessary delay, or needlessly increase the cost of litigation[, and is] neither unreasonable nor unduly burdensome or expensive, considering the needs of the case, prior discovery in the case, the amount in controversy, and the importance of the issues at stake in the action."[6] Similarly, Rule 26(b)(2)(C)(iii) requires the court to limit discovery when it determines that "the burden or expense of the proposed discovery outweighs its likely benefit, considering the needs of the case, the amount in controversy, the parties' resources, the importance of the issues at stake in the action, and the importance of the dis-

specific technologies used in the processes that are the subjects of this study are detailed *infra* in Sections 3.1 and 3.2.

[4] Bruce Hedin et al., Conference Report, *Overview of the TREC 2009 Legal Track*, SP 500-278 NIST Special Publ'n: 18th Text REtrieval Conf. Proc. (2009), http://trec-legal.umiacs.umd.edu/LegalOverview09.pdf.; Douglas W. Oard et al., Conference Report, *Overview of the TREC 2008 Legal Track*, SP 500-277 NIST Special Publ'n: 17th Text REtrieval Conf. Proc. (2008),
http://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf.

[5] Hedin, *supra* note 4.

[6] Fed. R. Civ. P. 26(g)(1).

covery in resolving the issues."[7] Thus, the rules require that discovery requests and responses be *proportional*. Rule 37(a)(4), however, provides that an incomplete response must be treated as a failure to respond, and therefore, requires that discovery responses be *complete*.[8]

Together, these rules reflect the tension that exists in all e-discovery processes between completeness, on the one hand, and burden and cost on the other.[9] In assessing what e-discovery is reasonable and proportional, the parties and the court must balance these competing considerations. One of the greatest challenges facing legal stakeholders is determining whether or not the cost and burden of identifying and producing electronically stored information ("ESI") is commensurate with its importance in resolving the issues in dispute. In current practice, the problem of identifying responsive (or privileged) ESI, once it has been collected, is almost always addressed, at least in part, by a manual review process, the cost of which dominates the e-discovery process.[10] A natural question to ask, then, is whether this manual review process is the most effective and efficient one for identifying and producing the ESI most likely to resolve a dispute.

Our investigation addresses a fundamental uncertainty that arises in determining what is reasonable and proportional: Is it true that if a human examines every document from a particular source, that human will, as nearly as possible, correctly identify all and only the documents that should be produced? That is, does exhaustive manual review guarantee that production will be as complete and correct as possible? Or can technology-assisted review, in which a human examines only a fraction of the documents, do better?

2.1 Information Retrieval

The task of finding all and only the documents that meet some requirement is precisely one of information retrieval ("IR"), a subject of scholarly research for at

[7] Fed. R. Civ. P. 26(b)(2)(C)(iii).

[8] Fed. R. Civ. P. 37(a)(4).

[9] In addition, at the same time as the responding party seeks to produce *all* responsive documents, they also seek to identify *only* the responsive documents, in order to guard against overproduction or waiver of privilege. *See, e.g., Mt. Hawley Insur. Co.* v. *Felman Prod., Inc.*, 271 F.R.D. 125, 136 (S.D. W.V. 2010) (30% overproduction cited as a factor in waiver of privilege).

[10] Marisa Peacock, *The True Cost of eDiscovery*, CMSWire, http://www.cmswire.com/cms/enterprise-cms/the-true-cost-of-ediscovery-006060.php (2009) (citing Sedona, *supra* note 2, at 192); Ashish Prasad, Kim Leffert & Shauna Fulbright-Paxton, *Cutting to the "Document Review" Chase: Managing a Document Review in Litigation and Investigations*, 18:2 Business Law Today (2008).

least a century.[11] In IR terms, "some requirement" is referred to as an *information need*, and *relevance* is the property of whether or not a particular document meets the information need. For e-discovery, the information need is typically specified by a production request (or by the rules governing privilege), and the definition of relevance follows. Cast in IR terms, the objective of review in e-discovery is to identify as many of the *relevant* documents as possible, while at the same time identifying as few *nonrelevant* documents as possible. The fraction of relevant documents that are identified is known as *recall*, while the fraction of identified documents that are relevant is known as *precision*. That is, recall is a measure of *completeness*, while precision is a measure of *accuracy*, or *correctness*.

A review result that has higher recall and higher precision than another is more nearly complete and correct, and therefore superior. One with lower recall and lower precision is inferior. If one result has higher recall while the other has higher precision, it is not immediately obvious which should be considered superior. F_1 – the harmonic mean of recall and precision[12] – is a commonly used summary measure that rewards results that achieve both high recall and high precision, while penalizing those that have either low recall or low precision.[13] The value of F_1 is always intermediate between recall and precision; closer to the lesser of the two. For example, a result with 40% recall and 60% precision (or 60% precision and 40% recall) has $F_1 = 48\%$. Following TREC, we report recall and precision, along with F_1 as a summary measure of overall review effectiveness.[14]

The notion of *relevance*, although central to information science, and the subject of much philosophical and scientific investigation,[15] remains elusive. While it is easy enough to write a document describing an information need and hence relevance, determining the relevance of any particular document requires human interpretation. It is well established that human assessors will disagree in a substantial number of cases whether a document is relevant, regardless of the infor-

[11] The concepts and terminology outlined in section 2.1 may be found in many information retrieval text books. For a historical perspective, *see* Gerard Salton & Michael J. McGill, *Introduction to Modern Information Retrieval* (1983); Cornelis Joost van Rijsbergen, *Information Retrieval* (2d ed. 1979). For a more modern treatment, *see* Büttcher, *supra* note 3.

[12] $F_1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}}$.

[13] van Rijsbergen, *supra* note 11, at 112.

[14] Hedin, *supra* note 4, at 37.

[15] *See* Tefko Saracevic, Relevance: *A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part III: Behavior and Effects of Relevance*, 58:13 J. of the Am. Soc'y for Info. Sci. & Tech., 2126 (2007); Tefko Saracevic, Relevance: *A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II: Nature and manifestations of relevance*, 58:3 J. of the Am. Soc'y for Info. Sci. & Tech., 1915 (2007).

mation need or the expertise and diligence of the assessors.[16]

2.2 Assessor Overlap

The level of agreement between independent assessors may be quantified by *overlap* – also known as the *Jaccard index* – the number of documents identified as relevant by two independent assessors, divided by the number identified as relevant by either or both of them.[17] For example, suppose assessor A identifies documents {W,X,Y,Z} as relevant, while assessor B identifies documents {V,W,X}. Both assessors have identified two documents {W,X} as relevant, while one or both have identified five documents {V,W,X,Y,Z} as relevant. So the overlap is $\frac{2}{5} = 40\%$. Informally, overlap of less than 50% indicates that the assessors disagree on whether or not a document is relevant more often than they agree that one is relevant.

Voorhees[18] measured overlap between primary, secondary, and tertiary reviewers who each made 14,968 assessments of relevance for 13,435 documents,[19] with respect to 50 information needs (or "topics," in TREC parlance), in connection with the Ad Hoc Task of the Fourth Text Retrieval Conference ("TREC 4").[20] As illustrated in Table 1, the overlap between primary and secondary assessors was 42.1%; the overlap between primary and tertiary assessors was 49.4%; and the overlap between secondary and tertiary assessors was 42.6%. Each of the topics was carefully specified in writing by the primary assessor; all assessors were professional information retrieval experts. Perhaps due to the assessors' expertise, Voorhees' overlap results are among the highest reported for pairs of human assessors. In short, assessors disagree at least as often as they agree that a document is relevant. Voorhees further concludes:

[16] *See* Peter Bailey et al., *Relevance assessment: are judges exchangeable and does it matter?* SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval 667 (2008).

[17] *Id.* at 667; Ellen M. Voorhees, *Variations in relevance judgments and the measurement of retrieval effectiveness*, 36:5 Information Processing & Management 697, 701 (2000), http://www.cs.cornell.edu/courses/cs430/2006fa/cache/Trec_8.pdf; Raimundo Real & Juan M. Vargas, *The probabilistic basis of Jaccard's index of similarity*, 45:3 Syst. Biol. 380 (1996).

[18] Voorhees, *supra* note 17, at 697-716.

[19] Some of the documents were assessed for relevance to more than one topic.

[20] Donna Harman, *Overview of the Fourth Text REtrieval Conference (TREC 4)*, SP 500-236 NIST Special Publ'n: 4th Text REtrieval Conf. Proc. 1 (2004); Ellen M. Voorhees and Donna K. Harman, eds., *TREC - Experiment and Evaluation in Information Retrieval* (MIT Press 2005).

Assessment	Primary	Secondary	Tertiary
Primary	100%		
Secondary	42.1%	100%	
Tertiary	49.4%	42.6%	100%

Table 1: Overlap in relevance assessments by primary, secondary, and tertiary assessors for the TREC 4 Ad Hoc Task.

> *The scores for the [secondary and tertiary] judgments imply a practical upper bound on retrieval system performance is 65% precision at 65% recall since that is the level at which humans agree with one another.*[21]

It is not widely accepted that these findings apply to e-discovery. This "legal exceptionalism" appears to arise from common assumptions within the legal community:

1. that the information need (responsiveness or privilege) is more precisely defined for e-discovery than for classical information retrieval;

2. that lawyers are better able to assess relevance than the assessors employed for typical information retrieval tasks; and

3. that the most defensible way to ensure that a production is accurate is to have a lawyer examine each and every document.

Assumptions (1) and (2) are amenable to scientific inquiry, as is the overarching question of whether technology-assisted review can improve on exhaustive manual review. Assumption (3) – a legal opinion – should be informed by science.

Recently, Roitblat et al.[22] studied the level of agreement among review teams using data from the response to a Second Request produced to the Department of Justice ("DOJ") in connection with the acquisition of MCI by Verizon. A random sample of 5,000 documents was reviewed by two independent teams of professional assessors, Teams A and B. Roitblat et al. report the level of agreement and disagreement between the original production, Team A, and Team B, as a contingency matrix,[23] from which we calculated overlap, as shown in Table 2. The

[21] Voorhees, *supra* note 17, at 701.

[22] Herbert Roitblat, Anne Kershaw & Patrick Oot, *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. Am. Soc'y for Info. Sci. & Tech. 1 (2010).

[23] *Id.* at 74 (Table 1).

Assessment	Production	Team A	Team B
Production	100%		
Team A	16.3%	100%	
Team B	15.8%	28.1%	100%

Table 2: Overlap in relevance assessments between original production in a legal matter, and two subsequent manual reviews.

overlap between Team A and the original production was 16.3%; the overlap between Team B and the original production was 15.8%; and the overlap between Teams A and B was 28.1%. These and other studies of overlap indicate that relevance is not a concept that can be applied consistently by independent assessors, even if the information need is specified by a production request and the assessors are lawyers.

2.3 Assessor Accuracy

Measurements of overlap provide little information regarding the accuracy of particular assessors because there is no "gold standard" against which to compare them. One way to resolve this problem is to deem the judgments made by one assessor to be correct by definition, and to use that assessor's judgments as the gold standard for the purpose of evaluating the other assessor(s).

In Voorhees, the primary assessor composed the information need specification for each topic;[24] it may therefore be reasonable to take his or her assessments to be the gold standard. In Roitblat et al., a senior attorney familiar with the case adjudicated all cases of disagreement between Teams A and B.[25] Although Roitblat et al. sought to measure agreement,[26] it may be reasonable to use their "adjudicated results" as the gold standard. Their adjudicated results were determined by deeming a senior attorney's opinion to be correct in cases where Teams A and B disagreed, and by deeming the consensus to be correct in cases where they agreed.[27] Assuming these gold standards, Table 3 shows the effectiveness

[24] Voorhees, *supra* note 17, at 700.

[25] Roitblat, *supra* note 22, at 74.

[26] Roitblat, *supra* note 22, at 72 ("Formally, the present study is intended to examine the hypothesis: The rate of agreement between two independent reviewers of the same documents will be equal to or less than the agreement between a computer-aided system and the original review.").

[27] Roitblat *supra* note 22, at 74. ("The 1,487 documents on which Teams A and B disagreed

of the various assessors in terms of recall, precision, and F_1.[28] We see that recall ranges from 52.8% to 83.6%, while precision ranges from 55.5% to 81.9%, and F_1 ranges from 64.0% to 70.4%. All in all, these results appear to be reasonable, but hardly perfect. Can technology-assisted review improve on them?

2.4 Technology-Assisted Review Accuracy

In addition to the two manual review teams, Roitblat et al. had two service providers (Teams C and D) use technology-assisted review processes to classify each document in the dataset as relevant or not.[29] Unfortunately, the adjudicated results described in Section 2.3 were made available to one of the two service providers, and therefore, cannot be used as a gold standard to evaluate the accuracy of their efforts.[30]

Instead, Roitblat et al. report recall, precision, and F_1 using as an alternate gold standard the set of documents originally produced to, and accepted by, the DOJ.[31] There is little reason to believe that this original production, and hence the alternate gold standard, was perfect.[32] The first two rows of Table 4 show the recall and precision of manual review Teams A and B when evaluated with

---

were submitted to a senior Verizon litigator (P. Oot), who adjudicated between the two teams, again without knowledge of the specific decisions made about each document during the first review. This reviewer had knowledge of the specifics of the matter under review, but had not participated in the original review. This authoritative reviewer was charged with determining which of the two teams had made the correct decision.").

[28] Recall and precision for the secondary and tertiary assessors, using the primary assessor as the gold standard, are provided by Voorhees *supra* note 17, at 701 (Table 2); recall and precision for Teams A and B, using the adjudicated results as the gold standard, were derived from Roitblat et al. *supra* note 22, at 74 (Table 1); F_1 was calculated from recall and precision using the formula at *supra* note 12.

[29] Roitblat, *supra* note 22, at 75.

[30] Roitblat, *supra* note 22, at 74 ("One of these systems based its classifications in part on the adjudicated results of Teams A and B, but without any knowledge of how those teams' decisions were related to the decisions made by [the] original review team. As a result, it is not reasonable to compare the classifications of these two systems to the classifications of the two re-review teams, but it is reasonable to compare them to the classifications of the original review.").

[31] *Id.*

[32] Roitblat, *supra* note 22, at 76 ("The use of precision and recall implies the availability of a stable ground truth against which to compare the assessments. Given the known variability of human judgments, we do not believe that we have a solid enough foundation to claim that we know which documents are truly relevant and which are not. Nevertheless, in the interest of comparison with existing studies (e.g., TREC Legal 2008), Table 2 shows the computed precision and recall of each of the four assessments using the original review as its baseline.").

Study	Review	Recall	Precision	F_1
Voorhees	Secondary	52.8%	81.3%	64.0%
Voorhees	Tertiary	61.8%	81.9%	70.4%
Roitblat et al.	Team A	77.1%	60.9%	68.0%
Roitblat et al.	Team B	83.6%	55.5%	66.7%

Table 3: Precision, recall, and F_1 of manual assessments in studies by Voorhees, and Roitblat et al. Voorhees evaluated secondary and tertiary assessors with respect to a primary assessor, who was deemed to be correct. Roitblat et al. evaluated manual review Teams A and B by having a senior attorney adjudicate disagreements.

respect to this alternate gold standard.[33] These results are much worse than those reported in Table 3. Team A achieved 48.8% recall and 19.7% precision, while Team B achieved 52.9% recall and 18.3% precision. The corresponding F_1 scores were 28.1% and 27.2%, respectively – less than half of the F_1 scores achieved with respect to the gold standard derived using the senior attorney's opinion.

The recall and precision reported by Roitblat et al., computed using the original production as the gold standard, are dramatically different from those we computed using their adjudicated results as the gold standard. Nevertheless, both sets of results appear to suggest the *relative* accuracy between Teams A and B: Team B has higher recall, while Team A has higher precision and higher F_1, regardless of which gold standard is applied.

The last two rows of Table 4 show the effectiveness of the technology-assisted reviews conducted by teams C and D, as reported by Roitblat et al. using the original production as the gold standard. The results suggest that technology-assisted review Teams C and D achieve about the same recall as manual review Teams A and B, and somewhat better precision and F_1. However, due to the use of the alternate gold standard, the result is inconclusive.[34] Because the improvement of using technology-assisted review reported by Roitblat et al. is small compared to the difference between the results observed using the two different gold standards, it is difficult to determine whether the improvement represents a real difference in effectiveness as compared to manual review.

In a heavily cited study by Blair and Maron,[35] skilled paralegal searchers were

[33] *Id.*

[34] *Id.*

[35] David Blair & M. E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*, 28:3 Comm. of the ACM 225, 289 (1985).

Study	Review	Method	Recall	Precision	F_1
Roitblat et al.	Team A	Manual	48.8%	19.7%	28.1%
Roitblat et al.	Team B	Manual	52.9%	18.3%	27.2%
Roitblat et al.	Team C	Tech. Asst.	45.8%	27.1%	34.1%
Roitblat et al.	Team D	Tech. Asst.	52.7%	29.5%	37.8%

Table 4: Recall, precision, and F_1 of manual and technology-assisted review teams, evaluated with respect to the original production to the DOJ. The first two rows of this table differ from the last two rows of Table 3 only in the gold standard used for evaluation.

instructed to retrieve at least 75% of all documents relevant to 51 requests for information pertaining to a legal matter.[36] For each request, the searchers composed keyword searches using an interactive search system, retrieving and printing documents for further review.[37] This process was repeated until the searcher was satisfied that 75% of relevant documents had been retrieved.[38] Although the searchers believed they had found 75% of relevant documents, their average recall was only 20.0%.[39] On the other hand, the searchers achieved a high average precision of 79.0%.[40] From the published data,[41] we calculated the average F_1 score to be 28.0% – remarkably similar to to that observed by Roitblat et al. for manual review.[42]

Blair and Maron argue that the searchers would not have been able to achieve higher recall even if they had known there were many unretrieved relevant documents;[43] Salton disagrees.[44] He claims that it would have been possible for the searchers to achieve higher recall at the expense of lower precision, either by

[36] *Id.* at 291.

[37] *Id.*

[38] *Id.*

[39] *Id.* at 293; *see also* Maureen Dostert & Diane Kelly, *Users' Stopping Behaviors and Estimates of Recall*, SIGIR '09: Proceedings of the 32nd annual international ACM SIGIR conference on Research and development in information retrieval 820 (2009) (showing that most subjects in an interactive information retrieval experiment reported they had found about 51-60% of the relevant documents when, on average, recall was only 7%.)

[40] Blair, *supra* note 35, at 291.

[41] *Id.*

[42] Roitblat, *supra* note 22, at 76.

[43] Blair, *supra* note 35, at 295.

[44] Gerard Salton, *Another look at automatic text-retrieval systems*, 29:7 Comm. of the ACM 577, 648 (1986).

broadening their queries or by taking advantage of the *relevance ranking* capability of the search system.[45]

Overall, the literature offers us little reason to believe that manual review is perfect. But is it as complete and accurate as possible, or can it be improved upon by one or more technology-assisted approaches invented in the quarter century since Blair and Maron?

Recent results from TREC suggest that technology-assisted approaches may indeed be able to improve on manual review.[46] In the TREC 2008 Legal Track Interactive Task, H5, a San Francisco-based legal information retrieval firm,[47] employed a user-modeling approach[48] to achieve recall, precision, and F_1 of 62.4%, 81.0%, and 70.5%, respectively,[49] in response to a mock request to produce documents[50] from a 6,910,192-document collection[51] released under the tobacco Master Settlement Agreement.[52] In the course of this effort, H5 examined only 7,992 documents[53] – 860 times fewer than the 6,910,192 that would have been needed to be examined in an exhaustive manual review. Yet the results compare favorably with those previously reported for manual review or keyword search, exceeding what Voorhees characterizes as a "practical upper bound" on what may be achieved, given uncertainties in assessment.[54]

One of the authors was inspired to try to reproduce these results at TREC 2009 using an entirely different approach: statistical active learning, originally developed for e-mail spam filtering.[55] At the same time, H5 reprised their approach

---

[45] *Id.* at 649.

[46] *See* Hedin, *supra* note 4; Oard, *supra* note 4; .

[47] H5, http://www.h5.com (last visited Feb. 13, 2011).

[48] Christopher Hogan et al., *H5 at TREC 2008 Legal Interactive: User Modeling, Assessment & Measurement*, H5 (2008), http://trec.nist.gov/pubs/trec17/papers/h5.legal.rev.pdf (last visited Feb. 13, 2011).

[49] Oard, *supra* note 4, at 30.

[50] TREC 2008 Complaints and Production Requests: Complaint I, http://trec-legal.umiacs.umd.edu/topics/8I.pdf (last visited Feb. 19, 2011).

[51] ir@IIT Projects: CDIP, http://ir.iit.edu/projects/CDIP.html (last visited Feb. 13, 2011).

[52] Project Tobacco Settlement Agreement, Nat'l Assoc. of Attorneys Gen. (Nov. 1998), *available at* http://www.naag.org/backpages/naag/tobacco/msa/msa-pdf/MSA%20with%20Sig%20Pages%20and%20Exhibits.pdf (last visited Feb. 13, 2011).

[53] Hogan, *supra* note 48, at 8.

[54] Voorhees, *supra* note 17, at 701.

[55] Gordon V. Cormack & Mona Mojdeh, *Machine Learning for Information Retrieval: TREC 2009 Web, Relevance Feedback and Legal Tracks* (Univ. of Waterloo 2009), http://trec.nist.gov/pubs/trec18/papers/uwaterloo-cormack.WEB.RF.LEGAL.pdf (last visited Feb.

Topic	Production Request
201	All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in structured commodity transactions known as "prepay transactions."
202	All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in transactions that the Company characterized as compliant with FAS 140 (or its predecessor FAS 125).
203	All documents or communications that describe, discuss, refer to, report on, or relate to whether the Company had met, or could, would, or might meet its financial forecasts, models, projections, or plans at any time after January 1, 1999.
204	All documents or communications that describe, discuss, refer to, report on, or relate to any intentions, plans, efforts, or activities involving the alteration, destruction, retention, lack of retention, deletion, or shredding of documents or other evidence, whether in hard-copy or electronic form.
205	All documents or communications that describe, discuss, refer to, report on, or relate to energy schedules and bids, including but not limited to, estimates, forecasts, descriptions, characterizations, analyses, evaluations, projections, plans, and reports on the volume(s) or geographic location(s) of energy loads.
206	All documents or communications that describe, discuss, refer to, report on, or relate to any discussion(s), communication(s), or contact(s) with financial analyst(s), or with the firm(s) that employ them, regarding (i) the Company's financial condition, (ii) analysts' coverage of the Company and/or its financial condition, (iii) analysts' rating of the Company's stock, or (iv) the impact of an analyst's coverage of the Company on the business relationship between the Company and the firm that employs the analyst.
207	All documents or communications that describe, discuss, refer to, report on, or relate to fantasy football, gambling on football, and related activities, including but not limited to, football teams, football players, football games, football statistics, and football performance.

Table 5: Mock production requests ("Topics") composed for the TREC 2009 Legal Track Interactive Task.

Team	Topic	Reviewed	Produced	Recall	Precision	F_1
Waterloo	201	6,145	2,154	77.8%	91.2%	84.0%
Waterloo	202	12,646	8,746	67.3%	88.4%	76.4%
Waterloo	203	4,369	2,719	86.5%	69.2%	76.9%
H5	204	20,000	2,994	76.2%	84.4%	80.1%
Waterloo	207	34,446	23,252	76.1%	90.7%	82.8%
	Average:	15,521	7,973	76.7%	84.7%	80.0%

Table 6: Effectiveness of H5 and Waterloo submissions to the TREC 2009 Legal Track Interactive Task.

for TREC 2009.[56] The TREC 2009 Legal Track Interactive Task used the same design as TREC 2008, but used a different complaint,[57] and seven new mock requests to produce documents (*see* Table 5) from a new collection of 735,872 e-mail messages and attachments captured from Enron at the time of its collapse.[58] Each participating team was permitted to request as many topics as they wished, however, due to resource constraints, the most topics that any team was assigned was four of the seven.[59]

Together, H5 and Waterloo produced documents for five distinct TREC 2009 topics; the results of their efforts are summarized in Table 6. The five efforts employed technology-assisted processes, with the number of manually reviewed documents for each topic ranging from 4,369 to 34,446[60] (or 0.6% to 4.7% of the collection). That is, the total human effort for the technology-assisted processes – measured by the number of documents reviewed – was between 0.6% and 4.7% of that which would have been necessary for an exhaustive manual review of all 735,872 documents in the collection. The number of documents produced for each topic ranged from 2,154 to 23,252[61] (or 0.3% to 3.2% of the collection;

13, 2011).

[56] Hedin, *supra* note 4, at 6.

[57] TREC 2009 Legal Track – Complaint J (June 18, 2009), *available at* http://trec-legal.umiacs.umd.edu/LT09_Complaint_J_final.pdf (last visited Feb. 13, 2011).

[58] FERC: Industries – Information Released in Enron Investigation, http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp (last visited Feb. 13, 2011).

[59] Hedin, *supra* note 4, at 7; E-mail from Bruce Hedin to Gordon V. Cormack & Maura R. Grossman (Mar. 24, 2011 2:46 EDT) (on file with authors).

[60] Cormack, *supra* note 55, at 6; E-mail from Dan Brassil to Maura R. Grossman (Dec. 17, 2010, 15:21 EST) (on file with authors).

[61] Hedin, *supra* note 4, at 10.

about half the number of documents reviewed). Over the five efforts, the average recall and precision were 76.7% and 84.7%, respectively; no recall was lower than 67.3%, and no precision was lower than 69.2%, placing all five efforts above what Voorhees characterizes as a "practical upper bound" on what may be achieved, given uncertainties in assessment.[62]

Although it appears that the TREC results are better than those previously reported in the literature, either for manual or technology-assisted review, they do not include any direct comparison between manual and technology-assisted review. To draw any firm conclusion that one is superior to the other, one must compare manual and technology-assisted review efforts using the same information needs, the same dataset, and the same evaluation standard. Roitblat et al. is the only peer-reviewed study known to the authors suggesting that technology-assisted review *may be* superior to manual review – if only in terms of precision, and only by a small amount – based on a common information need, a common dataset, and a common gold standard, albeit one of questionable accuracy.[63]

The present study shows conclusively that the H5 and Waterloo efforts are superior to manual reviews conducted contemporaneously by TREC assessors, using the same topics, the same datasets, and the same gold standard. The manual reviews are those that were undertaken at the request of the TREC organizers for the purpose of evaluating the participating teams' submissions. In comparing the manual and technology-assisted review efforts, we use exactly the same adjudicated gold standard used at TREC.[64]

Section 3 details the TREC 2009 Legal Track Interactive Task, including the H5 and Waterloo efforts, as well as the TREC process for assessment and gold-standard creation. Section 4 uses statistical inference to compare the recall, precision, and F_1 results achieved by H5 and Waterloo with what would have been achieved by the TREC assessors, had they reviewed all 735,872 documents in the collection. Section 5 presents a qualitative analysis of the nature of manual review errors. Sections 6, 7, and 8, respectively, discuss the results, limitations, and conclusions associated with this study.

3 TREC Legal Track Interactive Task

TREC is an annual event hosted by NIST, with the following objectives:

[62] Voorhees, *supra* note 17, at 701.

[63] Roitblat, *supra* note 32.

[64] Hedin, *supra* note 4.

- to encourage research in information retrieval based on large test collections;
- to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.[65]

Since its inception in 2006,[66] the TREC Legal Track has had the goal:

- to develop search technology that meets the needs of lawyers to engage in effective discovery in digital document collections.[67]

Within the TREC Legal Track, the Interactive Task simulates the process of review of a large population of documents for responsiveness to one or more discovery requests in a civil litigation.[68] In 2008, the first year of the Interactive Task,[69] the population of documents that was used was the Illinois Institute of Technology Complex Document Information Processing ("IIT CDIP") Test Collection, version 1.0,[70] consisting of about seven million documents that were released in connection with various lawsuits filed against certain U.S. tobacco companies and affiliated research institutes.[71] A mock complaint, along with three associated requests for production, were composed for the purposes of the Interactive Task by lawyers familiar with tobacco-related litigation.[72] Participating teams were

[65] Text Retrieval Conference (TREC) Overview, http://trec.nist.gov/overview.html (last visited Feb. 13, 2011).

[66] Jason R. Baron, *The TREC Legal Track: Origins and Reflections on the First Year*, 8 Sedona Conf. J. 251, 253 (2007); Jason R. Baron, David D. Lewis & Douglas W. Oard, *TREC 2006 Legal Track Overview*, SP 500-272 NIST Special Publ'n: 15th Text REtrieval Conf. Proc. (2006), http://trec.nist.gov/pubs/trec15/papers/LEGAL06.OVERVIEW.pdf.

[67] Text Retrieval Conference (TREC) Tracks, http://trec.nist.gov/tracks.html (last visited Feb. 13, 2011).

[68] Oard, *supra* note 4, at 20.

[69] *Id.* at 2.

[70] ir@IIT Projects: CDIP, *supra* note 51.

[71] Oard, *supra* note 4.

[72] *Id.* at 3.

required to produce the responsive documents for one or more of the three requests.[73]

For TREC 2009, the population of documents consisted of e-mail messages, with attachments, produced by Enron in response to requests by the Federal Energy Regulatory Commission ("FERC").[74] A mock complaint, along with seven associated requests for production, were composed for the purposes of the TREC effort.[75] Participating teams were required to produce the responsive documents for up to four of the requests, as assigned by TREC.[76]

Aside from the document collections, the mock complaints, and the production requests, the conduct of the 2008 and 2009 Interactive Tasks was identical. Participating teams were given the document collection, the complaint, and the production requests several weeks before production was due. Teams were allowed to use any combination of technology and human input; the exact combination differed from team to team. However, the size of the document population, along with time and cost constraints, rendered it infeasible for any team to conduct an exhaustive review of every document. To our knowledge, no team examined more than a small percentage of the document population; H5 and Waterloo, in particular, used various combinations of computer search, knowledge engineering, machine learning, and sampling to select documents for manual review.[77]

To aid the teams in their efforts, as well as to render an authoritative interpretation of responsiveness (or relevance, within the context of TREC), a volunteer *Topic Authority* ("TA") – a senior attorney familiar with the subject matter – was assigned for each topic. The TA played three critical roles:

- to consult with the participating teams to clarify the notion of relevance, in a manner chosen by the teams;
- to prepare a set of written guidelines used by the human reviewers to evaluate, after the fact, the relevance of documents produced by the teams; and
- to act as a final arbiter of relevance in the evaluation process.[78]

[73] *Id.* at 24.
[74] FERC, *supra* note 58.
[75] Hedin, *supra* note 4, at 5.
[76] *Id.* at 7 (Table 1).
[77] Hogan, *supra* note 48; Cormack, *supra* note 55.
[78] *See* Hedin, *supr*a note 4, at 2; Oard, *supra* note 4, at 20.

The various participant efforts were evaluated using estimates of recall, precision, and F_1 based on a two-pass human assessment process.[79] In the first pass, a stratified sample of about 7,000 documents was assessed for relevance by a team of human reviewers.[80] For some topics (Topics 201, 202, 205, and 206), the team consisted primarily of volunteer law students supervised by the TREC coordinators; for others (Topics 203, 204, and 207), the team consisted of lawyers employed and supervised by professional review companies, who volunteered their services.[81]

The first pass assessments were released to participating teams, who were invited to appeal those relevance determinations with which they disagreed.[82] The appeals were adjudicated by the TA, whose opinion was deemed to be correct and final.[83] The *gold standard* of relevance for the documents in each sample was therefore:

- the same as the first pass assessment, for any document that was not appealed; or
- the opinion rendered by the TA, for any document that was appealed.

Statistical inference was used to estimate recall, precision, and F_1 for the results produced by each participating team.[84]

Assuming that the participants were diligent in appealing the first-pass assessments with which they disagreed, it is reasonable to conclude that TREC's two-pass assessment process yielded a reasonably accurate gold standard. Moreover, we can use that same gold standard to evaluate not only the participants' submissions, but also to evaluate the first-pass assessments rendered by the human reviewers.

Subsections 3.1 and 3.2 *infra* briefly describe the processes employed by the two participants whose results we compare to manual review. Notably, the methods used by these participants differ substantially from those typically described in the industry as "clustering" or "concept search."[85]

[79] Hedin, *supra* note 4, at 3.
[80] *Id.* at 12-14.
[81] *Id.* at 8.
[82] *Id.* at 3.
[83] *Id.*
[84] *Id.* at 11-18.
[85] Sedona, *supra* note 2, at 202-03.

3.1 H5 Participation

For TREC 2009, H5 was assigned one topic (Topic 204).[86] Paraphrasing Dan Brassil of H5, the H5 process involves three steps: (i) definition of relevance, (ii) partly-automated design of deterministic queries, and (iii) measurement of precision and recall.[87] Once relevance is defined, the remaining processes of sampling and query design, and measurement of precision and recall, are conducted iteratively – allowing for query refinement and correction – until the clients' accuracy requirements are met.[88]

H5 describes how its approach differs from other information retrieval methods as follows:

> "It utilizes an iterative issue-focusing and data-focusing methodology that defines relevancy in detail (see [Brassil et al.[89]] for a brief discussion of issue- and data-focusing strategies . . .); most alternative processes provide a reductionist view of relevance (e.g.: a traditional coding manual), or assume that different individuals share a common understanding of relevance. . . . [H5's approach] is deterministic: each document is assessed against the relevance criteria and a relevant / not relevant determination is made. . . . [The approach] is built on precision: whereas many alternative approaches start with a small number . . . of keywords intended to be broad so as to capture a lot of relevant data (with the consequence of many false positives), H5's approach is focused on developing in an automated or semi-automated fashion large numbers of deterministic queries that are very precise: each string may capture just a few documents, but nearly all documents so captured will be relevant; and all the strings together will capture most relevant documents in the collection."[90]

For TREC 2009, H5 sampled and reviewed 20,000 documents to serve as input to

[86] Hedin, *supra* 4, at 6-7.

[87] E-mail from Dan Brassil to Maura R. Grossman (Dec. 17, 2010, 15:21 EST) (on file with authors).

[88] *Id.*

[89] Dan Brassil, Christopher Hogan & Simon Attfield, *The centrality of user modeling to high recall with high precision search*, in the Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics, 91 (2009).

[90] Brassil, *supra* note 87; as amended by e-mail from Dan Brassil to Maura R. Grossman (Feb. 16, 2011, 15:58 EST) (on file with authors).

various aspects of its process/methodology.[91] H5 declined to quantify the number of person-hours they expended during the seven to eight week time period between the assignment of the topic and the final submission date.[92]

3.2 Waterloo Participation

The University of Waterloo ("Waterloo") was assigned four topics (Topics 201, 202, 203, and 207).[93] Waterloo's approach consisted of three phases: (i) interactive search and judging, (ii) active learning, and (iii) recall estimation.[94] The interactive search and judging phase used essentially the same tools and approach that was used by Waterloo at TREC 6.[95] The Wumpus search engine[96] was coupled to a custom web interface (*see* Figure 1) that showed document excerpts and permitted assessments to be coded with a single mouse click. The full e-mail message containing the document (including its attachments in their native form), was also available for reference. Over the four topics, about 11,000 documents were retrieved and reviewed, at an average rate of 3 documents per minute (or 20 seconds per document; 55 hours in total). The resulting assessments were used to train an on-line active learning system, previously developed for spam filtering.[97]

The active learning system yields an estimate of the probability that each document is relevant.[98] An efficient user interface was constructed for reviewing documents selected according to this relevance score (*see* Figure 2). The primary approach was to examine unjudged documents in decreasing order of score, skipping previously assessed documents. Each document was rendered as text and the reviewer hit a single key ("s" for relevant, or "h" for not relevant) to record the assessment and move on to the next document. Over the four topics, about 50,000 documents were reviewed, at an average rate of 20 documents per minute (3 sec-

[91] Brassil, *supra* note 87.

[92] E-mail from Dan Brassil to Maura R. Grossman (Dec. 17, 2010, 15:57 EST) (on file with authors); e-mail from Dan Brassil to Maura R. Grossman (Feb. 16, 2011, 15:58 EST) (on file with authors).

[93] Cormack, *supra* note 55, at 2.

[94] *Id*. at 2-3.

[95] Gordon V. Cormack et al., *Efficient construction of large test collections*, SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval 282, 284 (1998).

[96] *Wumpus File System Search*, Welcome to the Wumpus Search Engine, http://www.wumpus-search.org/ (last visited Feb. 13, 2011).

[97] Cormack, *supra* note 55, at 2-3.

[98] *Id.* at 5.

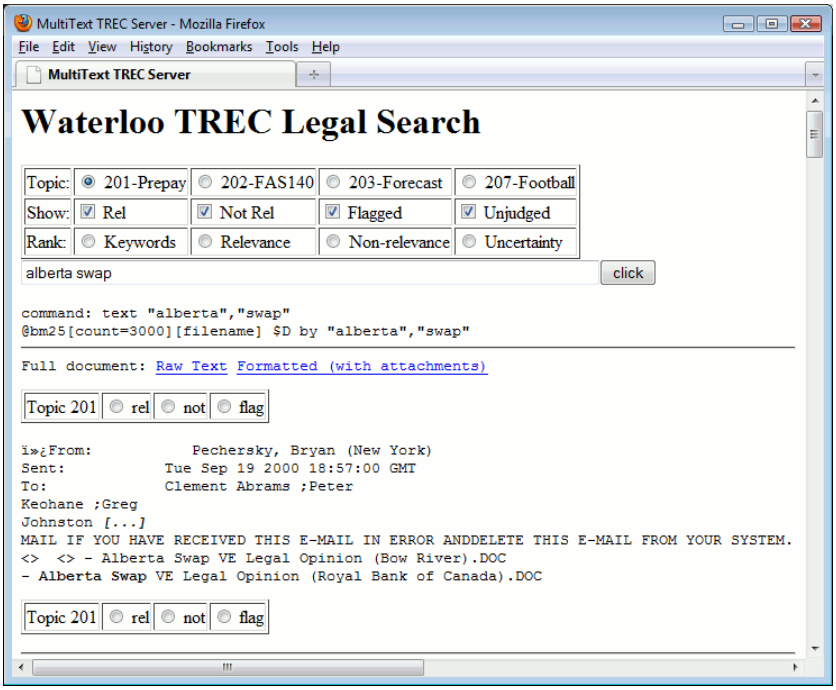

Figure 1: Interactive search and judging interface

onds per document; 42 hours in total).[99] From time to time, the interactive search and judging phase was revisited, to augment or correct the relevance assessments as new information came to light.[100]

The third and final phase is to estimate the density of relevant documents as a function of the score assigned by the active learning system, based on the assessments rendered during the active learning phase.[101] This estimate is used to gauge the tradeoff between recall and precision, and to determine the number of documents to produce so as to optimize F_1, as required by the task guidelines.[102]

For TREC 2009, the end result was that every document produced was reviewed by a human; however the number of documents reviewed was a small fraction of the entire document population (14,396 of 735,872 documents were reviewed, on average, per topic). Total review time for all phases was about 100 hours; 25 hours per topic, on average.

[99] *Id.* at 6.
[100] *Id.*
[101] *Id.*
[102] Hedin, *supra* 4, at 3.

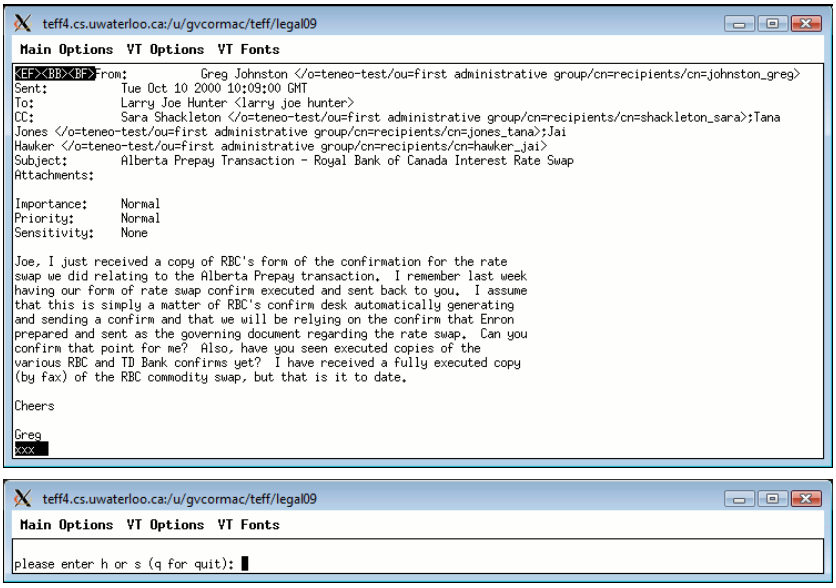

Figure 2: Minimalist review interface

4 Quantitative Analysis

Our aim is to refute the hypothesis that manual review is the best approach by showing that technology-assisted review can yield results that are more nearly complete and more accurate, as measured by recall, precision, and F_1. To compare technology-assisted to manual review, we require:

1. The results of one or more technology-assisted reviews. For this purpose, we use the review conducted by H5, and the four reviews conducted by Waterloo, in the course of their participation in the TREC 2009 Legal Track Interactive Task.[103]

2. The results of manual reviews for the same topics and datasets as the technology-assisted reviews. For this purpose, we use the manual reviews conducted by TREC on document samples for the purpose of evaluating the results submitted by the participating teams.[104]

[103] On file at the National Institute of Standards and Technology, available for research purposes subject to a usage agreement.

[104] Text Retrieval Conference (TREC) 2009 Legal Track Relevance Judgments and Evaluation Tools for the Interactive Task, http://trec.nist.gov/data/legal/09/evalInt09.zip (last visited Feb. 16, 2011).

3. A gold standard determination of relevance or nonrelevance.[105] For this purpose, we use the TREC final adjudicated assessments, for which the TREC TA was the ultimate arbiter.

The results of the technology-assisted and manual reviews were evaluated in exactly the same manner using the TREC methodology[106] and the TREC gold standard. To compare the effectiveness of the reviews, we report, for each topic:

1. Recall, precision, and F_1 for both the technology-assisted and manual reviews.

2. The *difference* in recall, the difference in precision, and the difference in F_1 between technology-assisted and manual review. A positive difference in some measure indicates that technology-assisted review is superior in that measure, while a negative difference indicates that it is inferior.

3. The *significance of the difference* for each measure, expressed as P.[107] Traditionally, $P < 0.05$ is interpreted to mean that the difference is statistically significant; $P > 0.1$ is interpreted to mean that the measured difference is not statistically significant. Smaller values of P imply stronger significance; $P < 0.001$ indicates overwhelming significance. P was computed by using 100 bootstrap samples of paired differences to estimate the standard error of measurement, and assuming a two-tailed normal distribution.[108]

Table 7 shows recall, precision, and F_1 for the technology-assisted and manual reviews for each of the five topics, as well as the overall average for the five technology-assisted reviews and the five manual reviews. For brevity, the difference in each measure is not shown, but may be computed easily from the table. For example, for Topic 201, the difference in recall between Waterloo and TREC is $77.8\% - 75.6\% = +2.2\%$. For each topic and each measure, the larger value is marked with either (*) or (†); (*) indicates that the measured difference is overwhelmingly significant ($P < 0.001$), while (†) indicates that it is not statistically significant ($P > 0.1$). All of the measured differences are either overwhelmingly significant or not statistically significant.

[105] *Id.*
[106] Hedin, *supra* note 4.
[107] Büttcher, *supra* note 3, at 426.
[108] *Id.* at 412-31.

Topic	Team	Recall	Precision	F_1
201	Waterloo	(†) 77.8%	(*) 91.2%	(*) 84.0%
	TREC (Law Students)	75.6%	5.0%	9.5%
202	Waterloo	67.3%	(*) 88.4%	(*) 76.4%
	TREC (Law Students)	(†) 79.9%	26.7%	40.0%
203	Waterloo	(*) 86.5%	(*) 69.2%	(*) 76.9%
	TREC (Professionals)	25.2%	12.5%	16.7%
204	H5	(*) 76.2%	(*) 84.4%	(*) 80.1%
	TREC (Professionals)	36.9%	25.5%	30.2%
207	Waterloo	76.1%	(†) 90.7%	82.8%
	TREC (Professionals)	(†) 79.0%	89.0%	(†) 83.7%
Avg.	H5/Waterloo	(†) 76.7%	(*) 84.7%	(*) 80.0%
	TREC	59.3%	31.7%	36.0%

Table 7: Effectiveness of TREC 2009 Legal Track technology-assisted methods (H5 and Waterloo) compared to exhaustive manual review (TREC). Results marked (*) are superior and overwhelmingly significant ($P < 0.001$). Results marked (†) are superior but not statistically significant ($P > 0.1$).

5 Qualitative Analysis

Our quantitative results (*see* Table 7) show the recall of manual review to vary from about 25% (Topic 203) to about 80% (Topic 202). That is, between 20% and 75% of all relevant documents are *missed* by human assessors. Is this shortfall the result of clerical error, a misinterpretation of relevance, or disagreement over marginal documents whose responsiveness is debatable? If the missed documents are marginal, the shortfall may be of little consequence; but if the missed documents are clearly responsive, production may be inadequate, and under Federal Rule of Civil Procedure 37(a)(4), such a production could constitute a failure to respond.[109]

To address this concern, we examined the documents that were coded as *non-responsive* to Topics 204 and 207 by the TREC assessors, but were coded as *responsive* by H5 or Waterloo, respectively, and adjudicated to be *responsive* by the TA. Recall from Table 5 that Topic 204 concerned shredding and destruction of documents, while Topic 207 concerned football and gambling. We chose these topics because they were more likely to be easily accessible to the reader, as op-

[109] *See* Fed. R. Civ. P. 37(a)(4).

```
Date: Tuesday, January 22, 2002 11:31:39 GMT
Subject:

I'm in.  I'll be shredding 'till 11am so I should
haveplenty of time to make it.
```

Figure 3: Topic 204 Inarguable error. This document was coded nonrelevant by a professional assessor, although it clearly pertains to document shredding, as specified in the production request.

posed to other topics that were more technical in nature. In addition, both of these topics were assessed by lawyers employed by professional review companies, using accepted practices for manual review.

For Topic 204, 160 of the 8,658 assessed documents were coded as nonresponsive by the manual assessors and responsive by H5 and the TA; for Topic 207, 51 of 7,129 assessed documents met these same criteria except that the responsive determination was made by Waterloo and the TA. From these numbers we extrapolate that 1,918 and 1,273 responsive documents from the collection (for Topics 204 and 207, respectively) would have been missed, had the entire collection been the subject of an exhaustive manual review.

For each of these documents, we used our subjective judgment to assess whether the document had been miscoded due to:

- *Inarguable error*: Under any reasonable interpretation of relevance, the document should have been coded as responsive, but was not. Possible reasons for such error include fatigue or inattention, overlooking part of the document, poor comprehension, or data entry mistakes in coding the document. For example, a document about "shredding" (*see* Figure 3) is responsive on its face to Topic 204; similarly "Fantasy Football" (*see* Figure 4) is responsive on its face to Topic 207.

- *Interpretive error*: Under some reasonable interpretation of relevance – but not the interpretation provided by the TA in the topic guidelines – the document might be considered nonresponsive. For example, an automated message stating in effect, "your mailbox is nearly full; please delete unwanted messages" (*see* Figure 5) might be construed by an assessor as nonresponsive to Topic 204, although the TA defined it as responsive. A message

```
From:   Bass, Eric
Sent:   Thursday, January 17, 2002 11:19 AM
To:     Lenhart, Matthew
Subject:        FFL Dues

You owe $80 for fantasy football.  When can you pay?
```

Figure 4: Topic 207 Inarguable error. This document was coded nonrelevant by a professional assessor, although it clearly pertains to fantasy football, as specified in the production request.

concerning children's football (*see* Figure 6) might be construed as nonresponsive to Topic 207, although the TA defined it as responsive.

- *Arguable error*: Reasonable, informed assessors might disagree or find it difficult to determine whether or not the document met the TA's conception of responsiveness. (*See* Figures 7 and 8.)

In making our determinations, we consulted the mock complaint,[110] as well as the topic-specific assessment guidelines that were prepared to memorialize the TA's conception of relevance and that were given to the human assessors for reference purposes.[111] Table 8 summarizes our findings: the vast majority of missed documents are attributable either to inarguable errors or to misinterpretations of the definition of relevance (interpretive error). Remarkably, we found only 4% of all errors to be arguable.

6 Results and Discussion

Tables 6 and 7 show that, by all measures, the average efficiency and effectiveness of the five technology-assisted reviews surpasses that of the five manual re-

[110] TREC 2009, *supra* note 57.

[111] TREC-2009 Legal Track – Interactive Task, Topic-Specific Guidelines – Topic 204, http://plg1.cs.uwaterloo.ca/trec-assess/TopicGuidelines_204_.pdf (last visited Feb. 13, 2011); TREC-2009 Legal Track – Interactive Task, Topic-Specific Guidelines – Topic 207, http://plg1.cs.uwaterloo.ca/trec-assess/TopicGuidelines_207_.pdf (last visited Feb. 13, 2011).

```
WARNING: Your mailbox is approaching the size limit

This warning is sent automatically to inform you that
your mailbox is approaching the maximum size limit.
Your mailbox size is currently 79094 KB.

Mailbox size limits:

        When your mailbox reaches 75000 KB you will
receive this message.To check the size of your mailbox:

  Right-click the mailbox (Outlook Today),
  Select Properties and click the Folder Size button.
  This method can be used on individual folders as well.

To make more space available, delete any items that are
no longer needed such as Sent Items and Journal entries.
```

Figure 5: Topic 204 Interpretive error. This automated message was coded nonresponsive by a professional assessor, although the TA construed such messages to be responsive to Topic 204.

```
Subject:        RE: Meet w/ Belden

I need to leave at 3:30 today to go to my stepson's
football game.   Unfortunately, I have a 2:00 and 3:00
meeting already.   Is this just a general catch-up
discussion?
```

Figure 6: Topic 207 Interpretive error. The assessor may have construed a children's league football game to be outside of the scope of "gambling on football." The TA deemed otherwise.

```
Subject: Original Guarantees
Just a followup note:
We are still unclear as to whether we should continue
to send original  incoming and outgoing guarantees to
Global Contracts (which is what we have  been doing
for about 4 years, since the Corp. Secretary kicked us
out of  using their vault on 48 for originals because
we had too many documents).  I  think it would be
good practice if Legal and Credit sent the originals
to  the same place, so we will be able to find them
when we want them.  So my  question to y'all is, do
you think we should send them to Global Contracts,  to
you, or directly the the 48th floor vault (if they let
us!).
```

Figure 7: Topic 204 Arguable error. This message concerns *where* to store particular documents, not specifically their destruction or retention. Reasonable, informed assessors might disagree as to its responsiveness, even knowing the TA's conception of relevance.

```
Subject:        RE: How good is Temptation Island 2
They have some cute guy lawyers this year-but I bet you
probably watch that manly Monday night Football.
```

Figure 8: Topic 207 Arguable error. This message mentions football, but not a *specific* football team, player, or game. Reasonable, informed assessors might disagree about whether or not it is responsive according to the TA's conception of relevance.

	Error Type			
Topic	Inarguable	Interpretive	Arguable	Total
204	98	56	6	160
207	39	11	1	51
Total	137	67	7	211
Fraction	65%	31%	4%	100%

Table 8: Number of responsive documents missed by human assessors, categorized by the nature of the error. 65% of missed documents are relevant on their face. 31% of missed documents are clearly relevant, when the topic-specific guidelines are considered. Only 4% of the missed documents, in the opinion of the authors, have debatable responsiveness.

views. The technology-assisted reviews require, on average, human review of only 2.1% of the documents, a nearly *fifty-fold savings* over exhaustive manual review. For F_1 and precision, the measured difference is overwhelmingly significant ($P < 0.001$); for recall the measured difference is not statistically significant ($P > 0.1$). These measurements provide strong evidence that the technology-assisted processes studied here yield better overall results, and better precision, in particular, than the TREC manual review process. The measurements also suggest that the technology-assisted processes may yield better recall, but the statistical evidence is insufficiently strong to support a firm conclusion to this effect.

It should be noted that the objective of TREC participants was to maximize F_1, not recall or precision, per se. It happens that they achieved, on average, higher precision. Had the participants considered recall to be more important, they might have traded off precision (and possibly F_1) for recall, by using a broader interpretation of relevance, or by adjusting a sensitivity parameter in their software.

Table 7 shows that, for four of the five topics, the technology-assisted methods achieve substantially higher F_1 scores, largely due to their high precision. Nonetheless, for a majority of the topics, the technology-assisted methods achieve higher recall as well; for two topics, substantially higher. For Topic 207, there is no meaningful difference in effectiveness between technology-assisted and manual review, for any of the three measures. *There is not one single measure for which manual review is significantly better than technology-assisted review.*

For three of the five topics (Topics 201, 202, and 204) our results show no significant difference in recall between technology-assisted and manual review. This result is perhaps not surprising, since the recall scores are all on the order of 70% – the best that might reasonably be achieved, given the level of agreement among

human assessors. Our results support the conclusion that technology-assisted review can achieve at least as high recall as manual review, and higher precision, at a fraction of the review effort, and hence, a fraction of the cost.

7 Limitations

The TREC effort uses a mock complaint and production requests composed by lawyers to be as realistic as possible.[112] The role of TA is intended to simulate that of a senior attorney overseeing a real document review.[113] The dataset consists of real e-mail messages captured within the context of an actual investigation.[114] These components of the study are perhaps as realistic as might reasonably be achieved outside of an actual legal setting. One possible limitation is that the Enron story, and the Enron dataset, are both well known, particularly since the Enron documents are frequently used in vendor product demonstrations. Both participants and TAs may have had prior knowledge of both that affected their strategies and assessments. In addition, there is a tremendous body of extrinsic information that may have influenced participants and assessors alike, including the results of the actual proceedings, commentaries,[115] books,[116] and even a popular movie.[117] It is unclear what effect, if any, this may have had on the results.

In general, the TREC teams were privy to less detailed guidance than the manual reviewers, placing the technology-assisted processes at a disadvantage. For example, Topic 202 requires the production of documents related to "transactions that the Company characterized as compliant with FAS 140." Participating teams were required to undertake research to identify the relevant transactions, as well as the names of the parties, counterparties, and entities involved. Manual reviewers, on the other hand, were given detailed guidelines specifying these elements.

The manual review was conducted on a stratified sample containing a higher proportion of relevant documents than the collection as a whole. Statistical infer-

[112] Hedin, *supra* note 4, at 2; Oard, *supra* note 4, at 3, 24.

[113] Hedin, *supra* note 4, at 2; Oard, *supra* note 4, at 20.

[114] Hedin, *supra* note 4, at 4.

[115] *See*, *e.g.*, John C. Coffee Jr., *What caused Enron?: A capsule social and economic history of the 1990's*, 89:2 Cornell L. Rev. 269 (2004); Paul M. Healy and Krishna G. Palepu, *The Fall of Enron*, 17:2 J. Econ. Persp. 3 (2003).

[116] *See*, *e.g.*, Bethany McLean and Peter Elkind, *The Smartest Guys in the Room: The Amazing Rise and Scandalous Fall of Enron* (2004); Loren Fox, *Enron: The rise and fall* (2003).

[117] *Enron: The Smartest Guys in the Room* (Magnolia Pictures 2005), http://www.imdb.com/title/tt1016268/ (last visited Mar. 7, 2011).

ence was used to evaluate the result of reviewing every document in the collection. Beyond the statistical uncertainty – which is quantified by the significance level P – there is uncertainty as to whether manual reviewers would have had the same error rate had they reviewed the entire collection. It is not unreasonable to think that, because the proportion of relevant documents would have been lower in the collection than it was in the sample, reviewer recall and precision might have been even lower, because reviewers would have tended to miss the needles in the haystacks due to fatigue, inattention, boredom, and related human factors. This sampling effect, combined with the greater guidance provided to the human assessors, may have resulted in an overestimate of the effectiveness of manual review, and thus, understated the results of this study.

The appeals process involves reconsideration – and potential reversal – *only* of manual coding decisions that are appealed by one or more participating teams, presumably because their results disagree with coding decisions made by the manual reviewers. The appeals process depends on participants exercising due diligence in identifying the assessments with which they disagree. While it appears that H5 and Waterloo exercised such diligence, it became apparent during the course of our analysis that they had overlooked a few assessor errors. These erroneous assessments were deemed to be correct under the gold standard, with the net effect of overstating the effectiveness of manual review, while understating the effectiveness of technology-assisted review. It is also likely that some documents were coded incorrectly by the manual review and also by every technology-assisted process, and not appealed. The impact of the resulting errors on the gold standard would be to overstate both recall and precision for the manual review, as well as for the technology-assisted review, with no net advantage to either.

We considered here only the results of two of the eleven teams participating in TREC 2009, because we thought they were most likely to demonstrate that technology-assisted review can improve on an exhaustive manual approach. We considered all submissions by these two teams, which happened to be the most effective submissions for five of the seven topics. We did not consider Topics 205 and 206, because neither H5 nor Waterloo submitted results for them. Furthermore, due to a dearth of appeals, there is no reliable gold standard for Topic 206.[118] We were aware before conducting this analysis that the H5 and Waterloo submissions were the most effective for their respective topics. To show that

[118] Hedin, *supra* note 4, at 18 ("Topic 206 represents the one topic, out of the seven featured in the 2009 exercise, for which we believe the post-adjudication results are not reliable. . . . We do not believe, therefore, that any valid conclusions can be drawn from the scores recorded for this topic.").

the results are significant in spite of this prior knowledge, we apply Bonferroni correction,[119] which multiplies P by 11, the number of participating teams. Even under Bonferroni correction, the results are overwhelmingly significant.

8 Conclusions

Overall, the myth that exhaustive manual review is the most effective – and therefore, the most defensible – approach to document review is strongly refuted. Technology-assisted review can (and does) yield more accurate results than exhaustive manual review, with much lower effort. Of course, not all technology-assisted reviews (and not all manual reviews) are created equal. The particular processes found to be superior in this study are both interactive, employing a combination of computer and human input. While these processes require the review of orders of magnitude fewer documents than exhaustive manual review, neither entails the naïve application of technology absent human judgment. Future work may address *which* technology-assisted review process(es) will improve *most* on manual review, not *whether* technology-assisted review *can* improve on manual review.

Acknowledgments

Cormack's research is supported by the Natural Sciences and Engineering Research Council (Canada). The authors would like to thank Ellen Voorhees and Ian Soboroff at NIST for providing access to the raw TREC 2009 data. The authors gratefully acknowledge the helpful comments received from Hon. John M. Facciola (D.D.C.), Hon. Paul W. Grimm (D. Md.), and Hon. Andrew J. Peck (S.D.N.Y.) on an earlier draft of this paper.

[119] Büttcher, *supra* note 3, at 428.