DUKE LAW

CENTER FOR JUDICIAL STUDIES
JOHN K. RABIEJ, DIRECTOR
DUKE UNIVERSITY SCHOOL OF LAW
210 SCIENCE DRIVE • BOX 90362
DURHAM, NC 27708-0362
PH: (919) 613-7059 • FAX: (919) 613-7158
JUDICIALSTUDIES@LAW.DUKE.EDU
WWW.LAW.DUKE.EDU/JUDICIALSTUDIES

October 17, 2013

Honorable David G. Campbell
Chair, Advisory Committee on Civil Rules
Washington, DC 20544

Dear Judge Campbell:

At the Duke Law Conference on Technology-Assisted Review ("TAR"), held in Washington, DC, in May 2013, judges, practitioners, academics, and technology experts gathered to consider the use of TAR as a means of effectively conducting e-discovery at substantially reduced cost. A consensus was reached that advanced analytical software applications and other technologies that can screen for relevant and privileged documents, properly applied, are effective and cost-efficient e-discovery tools, in appropriate cases, but parties are reluctant to use them absent official recognition. Accordingly, the undersigned propose that the following sentence be added at the end of the first paragraph of the Committee Note to Rule 26(b)(1):

> As part of the proportionality considerations, parties are encouraged, in appropriate cases, to consider the use of advanced analytical software applications and other technologies that can screen for relevant and privileged documents in ways that are at least as accurate as manual review, at far less cost.[1]

Consensus Justification of Revised Committee Note

The proposed Committee Note revision is consistent with and supports the main thrust of the Rule 26 amendments recommended by the Committee: to reduce mounting e-discovery costs that continue to plague litigants. The 2012 RAND study concluded that the use of TAR, in appropriate cases, could save litigants as much as 75% of the e-discovery costs in an individual case.[2] RAND's findings are consistent with other scholarly studies on the subject.[3] Yet litigants continue to shy away from using TAR out of fear that the courts will reject its validity, resulting in expensive "do-overs" or sanctions.

The proposed draft is limited and carefully crafted. It is non-controversial. It does not require that TAR be used in any given case. Nor does it assert that all forms of TAR – or indeed, any particular

[1] An excerpt of the Committee Note as revised is attached as Appendix A.
[2] RAND study attached as Appendix B.
[3] Additional studies supporting RAND's findings attached as Appendix C.

form of TAR – are effective. Because it does not establish any new obligation or right, it properly belongs in the Committee Note as useful guidance in applying the Rule's proportionality considerations.

What this single sentence would accomplish – which is crucial – is to show that the Civil Rules Committee finds nothing inherently wrong with the use of advanced analytical software applications and other technologies that can screen for relevant and privileged documents, in fact, supports their use in appropriate cases. The proposed statement would go far in raising awareness of the potential benefits of TAR and in addressing litigants' fear that judges might reject or impose onerous conditions on the use of TAR. These goals may be apparent, but the fact remains that many lawyers have shied away from using TAR precisely because of such lack of awareness and fear. At a time when e-discovery costs continue to burgeon, is there any principled reason not to recognize advanced analytical software applications and other technologies that can screen for relevant and privileged documents in the Committee Note?

Impact of Committee Note Revision

Litigants with substantial financial resources are just beginning to use TAR, sometimes in conjunction with other e-discovery tools and approaches, but its adoption has been slow. And the number of litigants − especially those with limited financial resources – who do not use TAR remains high. The Committee's recognition of TAR as a useful tool, in appropriate cases, in this single sentence would go far in facilitating its use.

Absent official recognition, the use of advanced analytical software applications and other technologies that can screen for relevant and privileged documents will continue to move forward haltingly – with only well-heeled litigants able to take on the perceived risks – adding to wasteful e-discovery costs. The absence of official recognition also leaves open the possibility that an ill-founded opinion may be issued that would further retard the use of TAR, requiring years of rulemaking to undo the chilling impact, along the lines evidenced by the Committee's amendment of Rule 37.

Conclusion

The proposed revision is modest; it is non-controversial; its downside is minimal; yet the upside benefits of increasing awareness and reducing e-discovery costs are very real. Maintaining a lean, Spartan Committee Note is good rulemaking, but it should not come at the cost of moving the law forward.

Thank you for considering this recommendation.

Sincerely,

Maura R. Grossman
Of Counsel
Wachtell, Lipton, Rosen & Katz

Gordon V. Cormack
Professor
University of Waterloo

John K. Rabiej
Director
Duke Law Center for Judicial Studies

List of Individuals Supporting the Proposed Revision to the Committee Note

Honorable Andrew J. Peck
Magistrate Judge
U.S. District Court-Southern District of NY

Ian J. Wilson
CEO
Servient, Inc.

Laura M. Kibbe
Managing Director
Epiq Systems

Michael E. Klein
Assistant General Counsel
Altria Client Services, Inc.

Conor R. Crowley
Principal
Crowley Law Office

Danuta Panich
Shareholder
Ogletree, Deakins, Nash, Smoak, Stewart

Julia Brickell
Executive Director & General Counsel
H5

Anne Kershaw
Attorney & Founder
A. Kershaw, PC

Matt Miller
Senior Vice President
DiscoverReady, LLC

Henry J. Kelston
Senior Counsel
Milberg, LLP

Richard P. Perrin
E-Discovery Counsel
Dickstein Shapiro, LLP

Carolyn Southerland
Managing Director
Huron Legal

Honorable John M. Facciola
Magistrate Judge
U.S. District Court-District of Columbia

Karl Schieneman
President
Review Less, LLC

Douglas F. MacPhail, Esq.
Attorney
Legalpeople

Pearlynn G. Houck
Attorney
Robinson, Bradshaw & Hinson, PA

Constantin Aliferis
Director
NYU Center for Health Informatics

Thomas B. Metzloff
Professor of Law
Duke Law School

Daniel P. Brassil
Principal Consultant
H5

Elise Singer
Of Counsel
Fine Kaplan & Black

Jonathan K. Levine
Partner
Girard Gibbs, LLP

Heyward Bouknight III
Attorney
Robinson Bradshaw & Hinson

Jerone J. English, Esq.
Director of e-Discovery
Intel Corporation

Matthew Nelson, Esq.
eDiscovery Counsel
Symantec Corporation

Appendix A

Excerpt of the Committee Note as Revised

Committee Note

The scope of discovery is changed in several ways. Rule 26(b)(1) is revised to limit the scope of discovery to what is proportional to the needs of the case. The considerations that bear on proportionality are moved from present Rule 26(b)(2)(C)(iii). Although the considerations are familiar, and have measured the court's duty to limit the frequency or extent of discovery, the change incorporates them into the scope of discovery that must be observed by the parties without court order. As part of the proportionality considerations, parties are encouraged, in appropriate cases, to consider the use of advanced analytical software applications and other technologies that can screen for relevant and privileged documents in ways that are at least as accurate as manual review, at far less cost.

The amendment deletes the former provision authorizing the court, for good cause, to order discovery of any matter relevant to the subject matter involved in the action. Proportional discovery relevant to any party's claim or defense suffices. Such discovery may support amendment of the pleadings to add a new claim or defense that affects the scope of discovery.

The former provision for discovery of relevant but inadmissible information that appears reasonably calculated to lead to the discovery of admissible evidence is also amended. Discovery of nonprivileged information not admissible in evidence remains available so long as it is otherwise within the scope of discovery. Hearsay is a common illustration. The qualifying phrase — "if the discovery appears reasonably calculated to lead to the discovery of admissible evidence" — is omitted. Discovery of inadmissible information is limited to matter that is otherwise within the scope of discovery, namely that which is relevant to a party's claim or defense and proportional to the needs of the case. The discovery of inadmissible evidence should not extend beyond the permissible scope of discovery simply because it is "reasonably calculated" to lead to the discovery of admissible evidence.

Rule 26(b)(2)(A) is revised to reflect the addition of presumptive limits on the number of requests for admission under Rule 36. The court may alter these limits just as it may alter the presumptive limits set by Rules 30, 31, and 33.

Rule 26(b)(2)(C)(iii) is amended to reflect the transfer of the considerations that bear on proportionality to Rule 26(b)(1). The court still must limit the frequency or extent of proposed discovery, on motion or on its own, if it is outside the scope permitted by Rule 26(b)(1).

Appendix B

RAND Study (Link Below)

Where the Money Goes -- Understanding Litigant Expenditures for Producing Electronic Discovery

http://www.rand.org/content/dam/rand/pubs/monographs/2012/RAND_MG1208.pdf

Appendix C

Additional Studies

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

TECHNOLOGY-ASSISTED REVIEW IN E-DISCOVERY CAN BE MORE EFFECTIVE AND MORE EFFICIENT THAN EXHAUSTIVE MANUAL REVIEW

By Maura R. Grossman[*] & Gordon V. Cormack[†] [**]

Cite as: Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11 (2011), http://jolt.richmond.edu/v17i3/article11.pdf.

[*] Maura R. Grossman is counsel at Wachtell, Lipton, Rosen & Katz. She is co-chair of the E-Discovery Working Group advising the New York State Unified Court System, and a member of the Discovery Subcommittee of the Attorney Advisory Group to the Judicial Improvements Committee of the U.S. District Court for the Southern District of New York. Ms. Grossman is a coordinator of the Legal Track of the National Institute of Standards and Technology's Text Retrieval Conference ("TREC"), and an adjunct faculty member at Rutgers School of Law–Newark and Pace Law School. Ms. Grossman holds a J.D. from Georgetown University Law Center, and an M.A. and Ph.D. in Clinical/School Psychology from Adelphi University. The views expressed herein are solely those of the Author and should not be attributed to her firm or its clients.

[†] Gordon V. Cormack is a Professor at the David R. Cheriton School of Computer Science, and co-director of the Information Retrieval Group, at the University of Waterloo. He is a coordinator of the TREC Legal Track, and Program Committee member of TREC at large. Professor Cormack is the co-author of *Information Retrieval: Implementing and Evaluating Search Engines* (MIT Press, 2010), as well as more than 100 scholarly articles. Professor Cormack holds a B.Sc., M.Sc., and Ph.D. in Computer Science from the University of Manitoba.

[**] 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.

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. This Article offers 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. The evidence derives from an analysis of data collected from the TREC 2009 Legal Track Interactive Task, and shows 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.

I. INTRODUCTION

[1] *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery* 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.[1]

While the word *myth* suggests disbelief, literature on the subject contains little scientific evidence to support or refute the notion that automated methods, while improving on the efficiency of manual review, yield inferior results.[2] This Article presents evidence supporting the position that a technology-assisted process, in which humans examine only a small fraction of the document collection, can yield higher recall and/or precision than an exhaustive manual review process, in which humans code and examine the entire document collection.

[2] 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.[3] A human examines and

[1] The Sedona Conference, *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) [hereinafter *Sedona Search Commentary*].

[2] *Id.* at 194 ("The comparative efficacy of the results of manual review versus the results of alternative forms of automated methods of review remains very much an open matter of debate.").

[3] *See* Douglas W. Oard et al., *Evaluation of information retrieval for E-discovery,* 18:4 ARTIFICIAL INTELLIGENCE & LAW 347, 365 (2010) ("In some cases . . . the end user will interact directly with the system, specifying the query, reviewing results, modifying the

codes only those documents the computer identifies – a tiny fraction of the entire collection.[4] Using the results of this human review, the computer codes the remaining documents in the collection for responsiveness (or privilege).[5] 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.[6] In contrast, *exhaustive manual 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.[7]

[3] Relevant literature suggests that manual review is far from perfect.[8] Moreover, 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.[9] By analyzing data collected

query, and so on. In other cases, the end user's interaction with the system will be more indirect. . . .").

[4] *See Sedona Search Commentary supra* note 1, at 209.

[5] *See* Maura R. Grossman & Terry Sweeney, *What Lawyers Need to Know About Search Tools*, THE NAT'L L.J. (Aug. 23, 2010), *available at* http://www.law.com/jsp/lawtechnologynews/PubArticleLTN.jsp?id=1202470952987&slreturn=1&hbxlogin=1 ("'machine learning tools,' use 'seed sets' of documents previously identified as responsive or unresponsive to rank the remaining documents from most to least likely to be relevant, or to classify the documents as responsive or nonresponsive.").

[6] *See, e.g.*, *Sedona Search Commentary*, *supra* note 1, at 217–23; CORNELIS JOOST VAN RIJSBERGEN, INFORMATION RETRIEVAL 74-85 (2d ed. 1979). The specific technologies employed in the processes that are the subjects of this study are detailed *infra* Parts III.A. – III.B.

[7] *See, e.g.*, Herbert L. Roitblat et al., *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y. FOR INFO. SCI. AND TECH. 70, 70 (2010).

[8] *See, e.g.*, *Sedona Search Commentary*, *supra* note 1.

[9] Bruce Hedin et al., *Overview of the TREC 2009 Legal Track*, *in* NIST SPECIAL PUBLICATION: SP 500-278, THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2009) PROCEEDINGS 16 & tbl.5 (2009), *available at* http://trec-

during the course of the TREC 2009 Legal Track Interactive Task,[10] the Authors 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 document-review companies – had they conducted a manual review of the entire document collection.

[4] Part II of this Article describes document review and production in the context of civil litigation, defines commonly used terms in the field of information retrieval, and provides an overview of recent studies. Part III 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. Part IV uses statistical inference to compare the recall, precision, and F_1 scores that H5 and Waterloo achieved to those the TREC assessors would have achieved had they reviewed all of the documents in the collection. Part V presents a qualitative analysis of the nature of manual review errors. Parts VI, VII, and VIII, respectively, discuss the results, limitations, and conclusions associated with this study. Ultimately, this Article 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?

II. CONTEXT

[5] Under Federal Rule of Civil Procedure 26(g)(1) ("Rule 26(g)(1)"), an attorney of record must certify "to the best of [his or her] knowledge,

legal.umiacs.umd.edu/LegalOverview09.pdf; *see also* Douglas W. Oard et al., *Overview of the TREC 2008 Legal Track*, *in* NIST SPECIAL PUBLICATION: SP 500-277, THE SEVENTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2008) PROCEEDINGS 8 (2008), *available at* http://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf.

[10] *See* Hedin et al., *supra* note 9, at 2.

information, and belief formed after a reasonable inquiry," that every discovery request, response, or objection is

> consistent with [the Federal Rules of Civil Procedure] . . . not interposed for any improper purpose, such as to harass, 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."[11]

Similarly, Federal Rule of Civil Procedure 26(b)(2)(C)(iii) ("Rule 26(b)(2)(C)(iii)") requires a 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 discovery in resolving the issues."[12] Thus, Rules 26(g)(1) and 26(b)(2)(C)(iii) require that discovery requests and responses be *proportional*.[13] However, Federal Rule of Civil Procedure 37(a)(4) ("Rule 37(a)(4)") provides that "an evasive or incomplete disclosure, answer or response must be treated as a failure to disclose, answer, or respond[,]" and therefore requires that discovery responses be *complete*.[14] Together, Rules 26(g)(1), 26(b)(2)(C)(iii), and 37(a)(4) reflect the tension – between completeness on one hand, and burden and cost on the other – that exists in all electronic discovery ("e-discovery") processes.[15] In

[11] FED. R. CIV. P. 26(g)(1).

[12] FED. R. CIV. P. 26(b)(2)(C)(iii).

[13] The Sedona Conference, *The Sedona Conference Commentary on Proportionality in Electronic Discovery,* 11 SEDONA CONF. J. 289, 294 (2010) [hereinafter *Sedona Proportionality Commentary*].

[14] FED. R. CIV. P. 37(a)(4).

[15] Typically, a responding party will not only seek to produce *all* responsive documents, but to identify *only* the responsive documents, in order to guard against overproduction or waiver of privilege. *See*, *e.g.*, Mt. Hawley Ins. Co. v. Felman Prod., Inc., 271 F.R.D. 125, 136 (S.D.W. Va. 2010) (finding that plaintiff's over-production of documents by more than 30% was a factor in waiver of privilege).

assessing what is reasonable and proportional with respect to e-discovery, parties and courts must balance these competing considerations.[16]

[6] 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.[17] 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.[18] 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.

A. Information Retrieval

[7] The task of finding all, and only, the documents that meet "some requirement" is one of information retrieval ("IR"), a subject of scholarly

[16] *See* Harkabi v. Sandisk Corp., No. 08 Civ. 8203 (WHP), 2010 WL 3377338, at *1 (S.D.N.Y Aug. 23, 2010) ("Electronic discovery requires litigants to scour disparate data storage mediums and formats for potentially relevant documents. That undertaking involves dueling considerations: thoroughness and cost.").

[17] *See id.* at *8 ("Integral to a court's inherent power is the power to ensure that the game is worth the candle—that commercial litigation makes economic sense. Electronic discovery in this case has already put that principle in jeopardy."); Hopson v. Mayor of Balt., 232 F.R.D. 228, 232 (D. Md. 2005) ("This case vividly illustrates one of the most challenging aspects of discovery of electronically stored information—how properly to conduct Rule 34 discovery within a reasonable pretrial schedule, while concomitantly insuring that requesting parties receive appropriate discovery, and that producing parties are not subjected to production timetables that create unreasonable burden, expense, and risk of waiver of attorney-client privilege and work product protection"). *See generally Sedona Proportionality Commentary*, *supra* note 13.

[18] 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 Search Commentary*, *supra* note 1, at 192); Ashish Prasad et al., *Cutting to the "Document Review" Chase: Managing a Document Review in Litigation and Investigations*, 18 BUS. LAW TODAY, 2, Nov.–Dec. 2008.

research for at least a century.[19] 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.[20] 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.[21] Cast in IR terms, the objective of review in e-discovery is to identify as many *relevant* documents as possible, while simultaneously identifying as few *nonrelevant* documents as possible.[22] The fraction of relevant documents identified during a review is known as *recall*, while the fraction of identified documents that are relevant is known as *precision*.[23] That is, *recall* is a measure of completeness, while *precision* is a measure of accuracy, or correctness.[24]

[8] The notion of *relevance*, although central to information science, and the subject of much philosophical and scientific investigation, remains elusive.[25] While it is easy enough to write a document describing an

[19] The concepts and terminology outlined in Part II.A may be found in many information retrieval textbooks. For a historical perspective, see GERARD SALTON & MICHAEL J. MCGILL, INTRODUCTION TO MODERN INFORMATION RETRIEVAL (1983); VAN RIJSBERGEN, *supra* note 6. For a more modern treatment, see STEFAN BÜTTCHER ET AL., INFORMATION RETRIEVAL: IMPLEMENTING AND EVALUATING SEARCH ENGINES 33–75 (2010).

[20] *See* BÜTTCHER ET AL., *supra* note 19, at 5-6, 8.

[21] *See* Hedin et al., *supra* note 9, at 1.

[22] *See* VAN RIJSBERGEN, *supra* note 6, at 4.

[23] *See* David C. Blair & M. E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*, 28 COMMC'NS ACM 289, 290 (1985) ("Recall measures how well a system retrieves *all* the relevant documents; and Precision, how well the system retrieves *only* the relevant documents."); VAN RIJSBERGEN, *supra* note 6, at 112-13.

[24] *See* VAN RIJSBERGEN, *supra* note 6, at 113.

[25] *See* 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 J. AM. SOC'Y FOR INFO. SCI. & TECH. 1915 (2007); Tefko Saracevic, *Relevance: A Review of the Literature and a Framework for Thinking on the Notion in*

information need and hence relevance, determining the relevance of any particular document requires human interpretation.[26] It is well established that human assessors will disagree in a substantial number of cases as to whether a document is relevant, regardless of the information need or the assessors' expertise and diligence.[27]

[9] A review resulting in higher recall and higher precision than another review is more nearly complete and correct, and therefore superior,[28] while a review with lower recall and lower precision is inferior.[29] If one result has higher recall while the other has higher precision, it is not immediately obvious which should be considered superior. To calculate a review's effectiveness, researchers often employ F_1 – the harmonic mean of recall and precision[30] – a commonly used summary measure that rewards results achieving both high recall and high precision, while penalizing those that have either low recall or low precision.[31] The value of F_1 is always intermediate between recall and precision, but is generally closer to the lesser of the two.[32] For example, a result with 40% recall and 60% precision has F_1 = 48%. Following

Information Science. *Part III: Behavior and Effects of Relevance*, 58:13 J. AM. SOC'Y FOR INFO. SCI. & TECH. 2126 (2007).

[26] *See* Peter Bailey et al., *Relevance Assessment: Are Judges Exchangeable and Does It Matter?*, *in* SIGIR '08 PROCEEDINGS OF THE 31ST ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL 667 (2008); *see also* VAN RIJSBERGEN, *supra* note 6, at 112.

[27] *See* Bailey et al., *supra* note 26, at § 4.3.

[28] *See* Blair & Maron, *supra* note 23.

[29] *See id*.

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

[31] *See* BÜTTCHER ET AL., *supra* note 19, at 68.

[32] *See id.*

TREC, this Article reports recall and precision, along with F_1 as a summary measure of overall review effectiveness.[33]

B. Assessor Overlap

[10] 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 assessors.[34] 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 either or both have identified five documents {V,W,X,Y,Z} as relevant. So the overlap is 2/5, or forty percent. Informally, overlap of less than fifty percent indicates that the assessors disagree on whether or not a document is relevant more often than when they agree that a document is relevant.[35]

[11] In her study, *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, Ellen Voorhees measured overlap between primary, secondary, and tertiary reviewers who each made 14,968 assessments of relevance for 13,435 documents,[36] with respect to 49

[33] *See* Hedin et al., *supra* note 9, at 3.

[34] Ellen M. Voorhees, *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, 36 INFO. PROCESSING & MGMT 697, 700 (2000), *available at* http://www.cs.cornell.edu/courses/cs430/2006fa/cache/Trec_8.pdf ("Overlap is defined as the size of the intersection of the relevant document sets divided by the size of the union of the relevant document sets."); *see* CHRISTOPHER D. MANNING ET AL., AN INTRODUCTION TO INFORMATION RETRIEVAL 61 (2009) (draft), *available at* nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf; *see also* Raimundo Real & Juan M. Vargas, *The Probabilistic Basis of Jaccard's Index of Similarity*, 45 SYSTEMATIC BIOLOGY 380, 381 (1996).

[35] *See* Ellen M. Voorhees, *The Philosophy of Information Retrieval Evaluation*, *in* EVALUATION OF CROSS-LANGUAGE INFORMATION RETRIEVAL SYSTEMS SECOND WORKSHOP OF THE CROSS-LANGUAGE EVALUATION FORUM, CLEF 2001 DARMSTADT, GERMANY, SEPTEMBER 3-4, 2001 REVISED PAPERS 355, 364 (Carol Peters et al. eds., 2002).

[36] E-mail from Ellen M. Voorhees to Gordon V. Cormack (Jul. 31, 2019 14:34 EDT) (on file with authors). The numbers in the text are derived from the file,

information needs (or "topics," in TREC parlance), in connection with Ad Hoc Task of the Fourth Text Retrieval Conference ("TREC 4").[37] As illustrated in Table 1, the overlap between primary and secondary assessors was 42.1%;[38] the overlap between primary and tertiary assessors was 49.4%;[39] and the overlap between secondary and tertiary assessors was 42.6%.[40]

[12] Perhaps due to the assessors' expertise,[41] Voorhees' overlap results are among the highest reported for pairs of human assessors. Her findings demonstrate that assessors disagree at least as often as they agree that a document is relevant.[42] Voorhees concluded:

> 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.[43]

"threeWayJudgments," attached to Voorhees' e-mail. Some of the documents were assessed for relevance to more than one topic.

[37] Voorhees, *supra* note 34, at 708; *see also* Donna Harman, *Overview of the Fourth Text REtrieval Conference (TREC-4)*, *in* NIST SPECIAL PUBLICATION 500-236: THE FOURTH TEXT RETRIEVAL CONFERENCE (TREC-4) 2 (2004), *available at* http://trec.nist.gov/pubs/trec4/t4_proceedings.html (follow the first link under "PAPERS").

[38] *See infra* Table 1; *see also* Voorhees, *supra* note 34, at 701 tbl.1.

[39] *See infra* Table 1; *see also* Voorhees, *supra* note 34, at 701 tbl.1.

[40] *See infra* Table 1; *see also* Voorhees, *supra* note 34, at 701 tbl.1.

[41] All assessors were professional information retrieval experts. Voorhees, *supra* note 34, at 701.

[42] *See id*.

[43] *Id*.

[13] It is not widely accepted that these findings apply to e-discovery.[44] 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;[45]

2. that lawyers are better able to assess relevance and privilege than the non-lawyers typically employed for information retrieval tasks;[46] and

3. that the most defensible way to ensure that a production is accurate is to have a lawyer examine each and every document.[47]

[44] *See Sedona Search Commentary*, *supra* note 1 (noting the widespread perception that manual review is nearly perfect). If that perception were correct, manual reviewers would have close to 100% overlap, contrary to Voorhees' findings. Vorhees, *supra* note 34, at 701 tbl.1.

[45] Oard et al., *supra* note 3, at 362 ("It is important to recognize that the notion of relevance that is operative in E-discovery is, naturally, somewhat more focused than what has been studied in information seeking behavior studies generally . . . .").

[46] *Cf.* Alejandra P. Perez, *Assigning Non-Attorneys to First-Line Document Reviews Requires Safeguards*, THE E-DISCOVERY 4-1-1 (LeClairRyan), Jan. 2011, at 1, *available at* http://marketing.leclairryan.com/files/Uploads/Documents/the-e-discovery-4-1-1-01-21-2011.pdf (opining that non-attorney document reviewers typically require additional training, particularly regarding the legal concept of privilege).

[47] *See Sedona Search Commentary*, *supra* note 1, at 203 ("Some litigators continue to primarily rely upon manual review of information as part of their review process. Principal rationales [include] . . . the perception that there is a lack of scientific validity of search technologies necessary to defend against a court challenge . . . ."); *see also* Thomas E. Stevens & Wayne C. Matus, *A 'Comparative Advantage' To Cut E-Discovery Costs*, NAT'L L.J. (Sept. 4, 2008), http://www.law.com/jsp/nlj/PubArticle NLJ.jsp?id=1202424251053 (describing a "general reluctance by counsel to rely on anything but what they perceive to be the most defensible positions in electronic discovery, even if those solutions do not hold up any sort of honest analysis of cost or quality").

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

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.[48]

[14] Recently, Herbert Roitblat, Anne Kershaw, and Patrick Oot studied the level of agreement among review teams using data produced to the Department of Justice ("DOJ") in response to a Second Request that stemmed from MCI's acquisition of Verizon.[49] In their study, two independent teams of professional assessors, Teams A and B, reviewed a random sample of 5,000 documents.[50] Roitblat and his colleagues reported the level of agreement and disagreement between the original production, Team A, and Team B, as a contingency matrix,[51] from which the Authors calculated overlap, as shown in Table 2.[52] The overlap between Team A and the original production was 16.3%;[53] the overlap between Team B and the original production was 15.8%;[54] and the overlap between Teams A and B was 28.1%.[55] These and other studies of overlap

[48] Voorhees, *supra* note 34, at 701 tbl.1.

[49] *See* Roitblat et al., *supra* note 7, at 73.

[50] *See id*. at 73-74.

[51] *Id.* at 74 tbl.1.

[52] *See infra* Table 2.

[53] *Id*.

[54] *Id*.

[55] *Id*.

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.[56]

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 Second Request, and two subsequent manual reviews.[57]

C. Assessor Accuracy

[15] Measurements of overlap provide little information regarding the accuracy of particular assessors because there is no "gold standard" against which to compare them.[58] One way to resolve this problem is to deem one assessor's judgments correct by definition, and to use those judgments as the gold standard for the purpose of evaluating the other assessor(s).[59]

[16] In the Voorhees study, the primary assessor composed the information need specification for each topic.[60] It may therefore be reasonable to take the primary assessor's coding decisions to be the gold standard. In the Roitblat, Kershaw, and Oot study, a senior attorney familiar with the case adjudicated all instances of disagreement between Teams A and B.[61] Although Roitblat and his colleagues sought to

[56] *See* Roitblat et al., *supra* note 7, at 73; Voorhees, *supra* note 34.

[57] The Authors derived the information in Table 2 from the Roitblat, Kershaw, and Oot study. Roitblat et al., *supra* note 7, at 74; *see supra* para. 13.

[58] Roitblat et al., *supra* note 7, at 77.

[59] *See* Voorhees, *supra* note 34, at 700.

[60] *Id.*

[61] Roitblat et al., *supra* note 7, at 74.

measure agreement,[62] it may be reasonable to use their "adjudicated results" as the gold standard. These adjudicated results deemed the senior attorney's opinion correct in cases where Teams A and B disagreed, and deemed the consensus correct in cases where Teams A and B agreed.[63] Assuming these gold standards, Table 3 shows the effectiveness of the various assessors in terms of recall, precision, and F_1.[64] Note 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%.[65] All in all, these results appear to be reasonable, but hardly perfect. Can technology-assisted review improve on them?

D. Technology-Assisted Review Accuracy

[17] In addition to the two manual review groups, Roitblat, Kershaw, and Oot had two service providers (Teams C and D) use technology-assisted review processes to classify each document in the dataset as

[62] *Id.* 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*.").

[63] *Id.* at 74.

> The 1,487 documents on which Teams A and B disagreed 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.

Id.

[64] *See infra* Table 3. Recall and precision for the secondary and tertiary assessors, using the primary assessor as the gold standard, are provided by Voorhees, *supra* note 34, at 701 tbl.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 7, at 74 tbl.1; F_1 was calculated from recall and precision using the formula at *supra* note 30.

[65] *See infra* Table 3.

relevant or not.[66] Unfortunately, the adjudicated results described in Part II.C. were made available to one of the two service providers, and therefore, cannot be used as a gold standard to evaluate the accuracy of the providers' efforts.[67]

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: Recall, precision, 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 correct. The Authors computed recall, precision, and F_1 from the results reported by Roitblat et al. for Teams A and B, using their adjudicated results as the gold standard.[68]

[18] Instead, Roitblat and his colleagues reported recall, precision, and F_1 using, as an alternate gold standard, the set of documents originally produced to, and accepted by, the DOJ.[69] There is little reason to believe that this original production, and hence the alternate gold standard, was perfect.[70] The first two rows of Table 4 show the recall and precision of manual review Teams A and B when evaluated with respect to this

[66] Roitblat et al., *supra* note 7, at 74-75.

[67] *Id.* 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.").

[68] Voorhees, *supra* note 34, at 701 tbl.2; Roitblat et al. *supra* note 7, at 74 tbl.1.

[69] Roitblat et al., *supra* note 7, at 74.

[70] *Id.* 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.").

alternate gold standard.[71] These results are much worse than those in Table 3.[72] Team A achieved 48.8% recall and 19.7% precision, while Team B achieved 52.9% recall and 18.3% precision.[73] 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.[74]

[19] The recall and precision Roitblat, Kershaw, and Oot reported were computed using the original production as the gold standard, and are dramatically different from those shown in Table 3, which were computed using their adjudicated results as the gold standard.[75] 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.[76]

[20] 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, Kershaw, and Oot using the original production as the gold standard.[77] The results suggest that technology-assisted review Teams C and D achieved about the same recall as manual review Teams A and B, and somewhat better precision and F_1.[78] However, due to the use of the alternate gold standard, the result is inconclusive.[79] Because the

[71] *See id.* at 76 tbl.2; *infra* Table 4.

[72] *Compare supra* Table 3, *with infra* Table 4.

[73] *See infra* Table 4; *see also* Roitblat et al., *supra* note 7, at 74-76.

[74] *Compare supra* Table 3, *with infra* Table 4.

[75] *Compare supra* Table 3, *with infra* Table 4. *See generally* Roitblat et al., *supra* note 7, at 76 tbl.2.

[76] *See supra* Table 3; *infra* Table 4; Roitblat et al., *supra* note 7, at 76 tbl.2.

[77] *See infra* Table 4; *see also* Roitblat et al., *supra* note 7, at 74-75.

[78] *See infra* Table 4.

[79] *See* Roitblat et al., *supra* note 7, at 76 ("The use of precision and recall implies the availability of a stable ground truth against which to compare the assessments. Given the

improvement from using technology-assisted review, as reported by Roitblat and his colleagues, 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.

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.[80]

[21] In a heavily cited study by David C. Blair and M.E. Maron, skilled paralegal searchers were instructed to retrieve at least 75% of all documents relevant to 51 requests for information pertaining to a legal matter.[81] For each request, the searchers composed keyword searches using an interactive search system, retrieving and printing documents for further review.[82] This process was repeated until the searcher was satisfied that 75% of the relevant documents had been retrieved.[83] Although the searchers believed they had found 75% of the relevant documents, their average recall was only 20.0%.[84] Despite this low rate of

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.").

[80] *Id.* at 73-76.

[81] *See* Blair & Maron, *supra* note 23, at 291.

[82] *Id.*

[83] *Id.*

[84] *Id.* at 293; *see also* Maureen Dostert & Diane Kelly, *Users' Stopping Behaviors and Estimates of Recall*, *in* SIGIR '09 PROCEEDINGS OF THE 32ND ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL 820–21 (2009) (showing that most subjects in an interactive information

recall, the searchers achieved a high average precision of 79.0%.[85] From the published data,[86] the Authors calculated the average F_1 score to be 28.0% – remarkably similar to that observed by Roitblat and his colleagues for manual review.[87]

[22] Blair and Maron argue that the searchers would have been unable to achieve higher recall even if they had known there were many relevant documents that were not retrieved.[88] Researcher Gerald Salton disagrees.[89] He claims that it would have been possible for the searchers to achieve higher recall at the expense of lower precision, either by broadening their queries or by taking advantage of the relevance ranking capability of the search system.[90]

[23] Overall, the literature offers 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 technology-assisted approaches invented since Blair and Maron's study?

[24] As previously noted, recent results from TREC suggest that technology-assisted approaches may indeed be able to improve on manual review.[91] In the TREC 2008 Legal Track Interactive Task, H5, a San

retrieval experiment reported they had found about 51-60% of the relevant documents when, on average, recall was only 7%).

[85] *See* Blair & Maron, *supra* note 23, at 293.

[86] *Id*.

[87] *See* Roitblat et al., *supra* note 7 at 76.

[88] *See* Blair & Maron, *supra* note 23, at 295-96.

[89] *See* Gerard Salton, *Another Look at Automatic Text-Retrieval Systems*, 29:7 COMMC'NS ACM 648, 650 (1986).

[90] *Id.* at 648-49.

[91] *See generally* Hedin et al., *supra* note 9; Oard et al., *supra* note 9.

Francisco-based legal information retrieval firm,[92] employed a user-modeling approach[93] to achieve recall, precision, and F_1 of 62.4%, 81.0%, and 70.5%, respectively, in response to a mock request to produce documents from a 6,910,192-document collection released under the tobacco Master Settlement Agreement.[94] In the course of this effort, H5 examined only 7,992 documents[95] – roughly 860 times fewer than the 6,910,192 it would have been necessary to examine 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.[96]

[92] *See Contact Us*, H5, http://www.h5.com/about/contact.php (last visited Mar. 22, 2011); *Who We Are*, H5, http://www.h5.com/about/who_we_are.html (last visited Apr. 11, 2011).

[93] Christopher Hogan et al., *H5 at TREC 2008 Legal Interactive: User Modeling, Assessment & Measurement*, *in* NIST SPECIAL PUBLICATION: SP 500-277, THE SEVENTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2008) PROCEEDINGS (2008), *available at* http://trec.nist.gov/pubs/trec17/papers/ h5.legal.rev.pdf (last visited Mar. 23, 2011).

[94] Oard et al., *supra* note 9, at 30 tbl.15; *see also Complex Document Image Processing (CDIP)*, ILL. INST. TECH., http://ir.iit.edu/projects/ CDIP.html (last visited Apr. 11, 2011); *Master Settlement Agreement*, NAT'L ASS'N ATTORNEYS GEN. (Nov. 1998), *available at* http://www.naag.org/backpages/naag/tobacco/msa/msa-pdf/MSA%20with%20Sig%20 Pages%20and%20Exhibits.pdf; TREC 2008, *Complaint for Violation of the Federal Securities Laws, Mellon v. Echinoderm Cigarettes, Inc.*, (2008), *available at* http://trec-legal.umiacs.umd.edu/topics/8I.pdf.

[95] Hogan et al., *supra* note 92, at 8.

[96] Voorhees, *supra* note 34, at 701.

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.[97]

[97] TREC 2009, *Complaint, Grumby v. Volteron Corp.*, 14 (2009) *available at* http://*trec-legal.umiacs.umd.edu/LT09_Complaint _J_final.pdf*; *see also* Hedin et al., *supra* note 9, at 5-6.

[25] 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.[98] At the same time, H5 reprised its approach for TREC 2009.[99] The TREC 2009 Legal Track Interactive Task used the same design as TREC 2008, but employed a different complaint[100] and seven new mock requests to produce documents (see Table 5) from a new collection of 836,165 e-mail messages and attachments captured from Enron at the time of its collapse.[101] 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.[102]

[98] *See generally* Gordon V. Cormack & Mona Mojdeh, *Machine Learning for Information Retrieval: TREC 2009 Web, Relevance Feedback and Legal Tracks*, *in* NIST SPECIAL PUBLICATION: SP 500-278, THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2009) PROCEEDINGS (2009), *available at* http://trec.nist.gov/pubs/trec18/papers/uwaterloo-cormack.WEB.RF.LEGAL.pdf.

[99] Hedin et al., *supra* note 9, at 6.

[100] *See generally* TREC 2009, *Complaint*, *supra* note 97.

[101] Hedin et al., *supra* note 9, at 4; *see Information Released in Enron Investigation*, FED. ENERGY REG. COMM'N, http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp (last visited Apr. 11, 2011) [hereinafter FERC]; E-mail from Bruce Hedin to Gordon V. Cormack (Aug. 31, 2009 20:33 EDT) (on file with authors) ("I have attached full list of the 836,165 document-level IDs . . . ."). The collection is available at *Practice Topic and Assessments for TREC 2010 Legal Learning Task*, U. WATERLOO, http://plg1.uwaterloo.ca/~gvcormac/treclegal09/ (follow "The TREC 2009 dataset") (last visited Apr. 18, 2011).

[102] Hedin et al., *supra* note 9, at 7; E-mail from Bruce Hedin to Gordon V. Cormack & Maura R. Grossman (Mar. 24, 2011 02:46 EDT) (on file with authors).

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.[103]

[26] Together, H5 and Waterloo produced documents for five distinct TREC 2009 topics;[104] 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[105] (or 0.5% to 4.1% of the collection). That is, the total human effort for the technology-assisted processes – measured by the number of documents reviewed – was between 0.5% and 4.1% of that which would have been necessary for an exhaustive manual review of all 836,165 documents in the collection.[106] The number of documents produced for each topic ranged from 2,154 to 23,252[107] (or 0.3% to 2.8% of the collection; about half the number of documents reviewed). Over the five efforts, the average recall and precision were 76.7% and 84.7%,

[103] *See infra*, para. 25.

[104] *See* Hedin et al., *supra* note 9, at 7.

[105] Cormack & Mojdeh, *supra* note 98, at 6 tbl.2 (showing that Waterloo reviewed between 4,369 documents (for Topic 203) and 34,446 documents (for Topic 207); *see* E-mail from Dan Brassil to Maura R. Grossman (Dec. 17, 2010 15:21 EST) (on file with authors) ("[H5] sampled and reviewed 20,000 documents").

[106] *See* sources cited *supra* note 101.

[107] NIST Special Publication 500-277: The Seventeenth Text REtrieval Conference Proceedings (TREC 2008) http://trec.nist.gov/pubs/trec17/t17_proceedings.html Appendix: Per Topic Scores: TREC 2009 Legal Track, Interactive Task, 3 tbl.4, 4 tbl.8, 5 tbl.12, 6 tbl.16, 9 tbl.26 http://trec.nist.gov/pubs/trec18/appendices/ app09int2.pdf.

respectively; no recall was lower than 67.3%, and no precision was lower than 69.2%,[108] placing all five efforts above what Voorhees characterized as a "practical upper bound" on what may be achieved, given uncertainties in assessment.[109]

[27] 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.[110] 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.[111] The Roitblat, Kershaw, and Oot study 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.[112]

[28] This Article 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 considered for this Article were the "First-Pass Assessments" undertaken at the request of the TREC coordinators for

[108] *See* Hedin et al, *supra* note 9, at 17.

[109] Voorhees, *supra* note 34, at 701.

[110] *See e.g.*, Oard et al., *supra* note 9, at 1-2.

[111] *See* Voorhees, *supra* note 35, at 356 ("The [Cranfield] experimental design called for the same set of documents and same set of information needs to be used for each [search method], and for the use of both precision and recall to evaluate the effectiveness of the search.").

[112] *See* Roitblat et al., *supra* note 7, 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.").

the purpose of evaluating the participating teams' submissions.[113] In comparing the manual and technology-assisted reviews, the Authors used exactly the same adjudicated gold standard as TREC.[114]

III. TREC Legal Track Interactive Task

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

- 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.[115]

Since its inception in 2006,[116] 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."[117]

[113] Hedin et al., *supra* note 9, at 3 (describing the "First-Pass Assessment" process).

[114] *See id*. at 3-4.

[115] Text REtrieval Conference (TREC), *Overview*, NAT'L INST. STANDARDS & TECH., http://trec.nist.gov/overview.html (last updated Aug. 10, 2010).

[116] *See* Jason R. Baron, *The TREC Legal Track: Origins and Reflections on the First Year*, 8 SEDONA CONF. J. 251, 253 (2007); *see also* Jason R. Baron et al., *TREC-2006 Legal Track Overview*, *in* NIST SPECIAL PUBLICATION: SP 500-272, THE FIFTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2006) PROCEEDINGS 1-2 (2006), *available at* http://trec.nist.gov/pubs/trec15/papers/LEGAL06.OVERVIEW.pdf.

[30] 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.[118] In 2008, the first year of the Interactive Task,[119] the population of documents used was the "Illinois Institute of Technology Complex Document Information Processing Test Collection, version 1.0" ("IIT CDIP"),[120] 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.[121] A mock complaint and three associated requests for production (or topics) were composed for the purposes of the Interactive Task.[122] Participating teams were required to produce the responsive documents for one or more of the three requests.[123]

[31] The population of documents used for TREC 2009 consisted of e-mail messages and attachments that Enron produced in response to requests by FERC.[124] A mock complaint and seven associated requests for production were composed for the purposes of TREC 2009.[125] Participating teams requested as many topics as they desired to undertake, but time and cost constraints limited the number of topics that any team was assigned to a maximum of four.[126]

[117] Text Retrieval Conference (TREC), *TREC Tracks*, NAT'L INST. STANDARDS & TECH., http://trec.nist.gov/tracks.html (last updated Feb. 24, 2011).

[118] *See* Oard et al., *supra* note 9, at 20.

[119] *See id.* at 2.

[120] *Id.* at 3; *see Complex Document Image Processing (CDIP)*, *supra* note 94.

[121] *See* Oard et al., *supra* note 9, at 3; *Complex Document Image Processing (CDIP)*, *supra* note 93.

[122] *See* Oard et al., *supra* note 9 at 3, 24.

[123] *Id.* at 24.

[124] *See* Hedin et al., *supra* note 9, at 4; *see also* FERC, *supra* note 101.

[125] *See* Hedin et al., *supra* note 9, at 5-6.

[126] *See id.* at 7 tbl.1.

[32] Aside from the document collections, the mock complaints, and the production requests, the conduct of the 2008 and 2009 Interactive Tasks was identical.[127] Participating teams were given the document collection, the complaint, and the production requests several weeks before production was due.[128] Teams were allowed to use any combination of technology and human input; the exact combination differed from team to team.[129] 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.[130] To the Authors' 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.[131]

[33] 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.[132] 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;

[127] *See id.* at 1-2.

[128] *See* Text Retrieval Conference (TREC), *TREC-2008 Legal Track Interactive Task: Guidelines*, 8, 17 (2008), trec-legal.umiacs.umd.edu/2008InteractiveGuidelines.pdf [hereinafter *TREC-2008 Guidelines*]; *see also* E-mail from Dan Brassil to Maura R. Grossman, *supra* note 105.

[129] *TREC-2008 Guidelines*, *supra* note 128, at 4, 7; *see also* E-mail from Bruce Hedin to Gordon V. Cormack (Apr. 07, 2011 00:56 EDT) (confirming that teams were permitted to use any combination of technology and human input).

[130] *See TREC-2008 Legal Track Interactive Task: Guidelines*, *supra* note 128, at 8.

[131] *See* Hogan et al., *supra* note 9, at 5; Cormack & Mojdeh, *supra* note 98, at 6.

[132] *See* Hedin et al., *supra* note 9, at 2.

- 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 adjudication process.[133]

[34] The TREC coordinators evaluated the various participant efforts using estimates of recall, precision, and F_1 based on a two-pass human assessment process.[134] In the first pass, human reviewers assessed a stratified sample of about 7,000 documents for relevance.[135] For some topics (Topics 201, 202, 205, and 206), the reviewers were primarily volunteer law students supervised by the TREC coordinators; for others (Topics 203, 204, and 207), the reviewers were lawyers employed and supervised by professional document-review companies, who volunteered their services.[136]

[35] The TREC coordinators released the first-pass assessments to participating teams, which were invited to appeal relevance determinations with which they disagreed.[137] For each topic, the TA adjudicated the appeals, and the TA's opinion was deemed to be correct and final.[138] The gold standard of relevance for the documents in each sample was therefore:

- The same as the first-pass assessment, for any document that participants did not appeal; or

[133] *Id.* at 2-3; *see* Oard et al., *supra* note 9, at 20.

[134] Hedin et al., *supra* note 9, at 3-4.

[135] *See id.* at 12-14.

[136] *Id.* at 8.

[137] *Id.* at 3.

[138] *Id.*

- The TA's opinion, for any document that participants did appeal.

The TREC coordinators used statistical inference to estimate recall, precision, and F_1 for the results each participating team produced.[139]

[36] Assuming participants diligently appealed the first-pass assessments with which they disagreed, it is reasonable to conclude that TREC's two-pass assessment process yields a reasonably accurate gold standard. Moreover, that same gold standard is suitable to evaluate not only the participants' submissions, but also the first-pass assessments of the human reviewers.[140]

[37] Parts III.A and III.B briefly describe the processes employed by the two participants whose results this Article compares to manual review. Notably, the methods the two participants used differ substantially from those typically described in the industry as "clustering" or "concept search."[141]

A. H5 Participation

[38] At TREC 2009, H5 completed one topic (Topic 204).[142] According to 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."[143] "Once relevance is defined, the two remaining processes of (1) sampling and query design and (2) measurement of precision and recall are conducted

[139] *Id.* at 3, 11-16.

[140] *See* Hedin et al., *supra* note 9, at 13 (describing the construction of the gold standard).

[141] *Sedona Search Commentary*, *supra* note 1, at 202-03.

[142] Hedin et al., *supra* note 9, at 6-7.

[143] E-mail from Dan Brassil to Maura R. Grossman, *supra* note 105.

iteratively – 'allowing for query refinement and correction' – until the clients' accuracy requirements are met."[144]

[39] 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; 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.[145]

In the course of its TREC 2009 effort, H5 sampled and reviewed a total of 20,000 documents.[146] H5 declined to quantify the number of person-hours

[144] *Id.*

[145] *Id.* (citing Dan Brassil et al., *The Centrality of User Modeling to High Recall with High Precision Search*, *in* 2009 IEEE Int'l Conf. on Systems, Man, and Cybernetics, 91, 91-96.

[146] *Id*.

it expended during the seven to eight week time period between the assignment of the topic and the final submission date.[147]

B. Waterloo Participation

[40] The University of Waterloo ("Waterloo") completed four topics (Topics 201, 202, 203, and 207).[148] Waterloo's approach consisted of three phases: (i) "interactive search and judging," (ii) "active learning," and (iii) recall estimation.[149] The interactive search and judging phase "used essentially the same tools and approach [Waterloo] used in TREC 6."[150] Waterloo coupled the Wumpus search engine[151] to a custom web interface that provided document excerpts and permitted assessments to be coded with a single mouse click.[152] Over the four topics, roughly 12,500 documents were retrieved and reviewed, at an average rate of about 3 documents per minute (about 22 seconds per document; 76 hours in

[147] *Id.*; E-mail from Dan Brassil to Maura R. Grossman (Feb. 16, 2011 15:58 EST) (on file with authors).

[148] Cormack & Mojdeh, *supra* 98, at 2.

[149] *Id.* at 1-3.

[150] *Id.* at 2. *See generally*, Gordon V. Cormack et al., *Efficient Construction of Large Test Collections*, *in* SIGIR '98 PROCEEDINGS OF THE 21ST ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL 282, 284 (1998).

[151] *Welcome to the Wumpus Search Engine!*, WUMPUS, http://www.wumpussearch.org/ (last visited Apr. 11, 2011).

[152] *See* Cormack & Mojdeh, *supra* note 98, at 3 & fig.2; *see also infra* Figure 1. "We used the Wumpus search engine and a custom html interface that showed hits-in-context and radio buttons for adjudication . . . . Available for reference were links to the full text of the document and to the full email message containing the document, including attachments in their native format." Cormack & Mojdeh, *supra* note 98, at 3.

total).[153] Waterloo used the resulting assessments to train an on-line active learning system, previously developed for spam filtering.[154]

[41] The active learning system "yields an estimate of the [probability] that each document is relevant."[155] Waterloo developed an "efficient user interface to review documents selected by this relevance score" (see Figure 2).[156] "The primary approach was to examine unjudged documents in decreasing order of score, skipping previously adjudicated documents."[157] The process displayed each document as text and, using a single keystroke, coded each document as relevant or not relevant.[158] Among the four topics, "[a]bout 50,000 documents were reviewed, at an average rate of 20 documents per minute (3 seconds per document)" or 42 hours in total.[159] "From time to time, [Waterloo] revisited the interactive search and judging system, to augment or correct the relevance assessments as new information came to light."[160]

[153] E-mail from Gordon V. Cormack to K. Krasnow Waterman (Feb. 24, 2010 08:25 EST) (on file with authors) (indicating that 12,508 documents were reviewed at a rate of 22 seconds per document, *i.e.*, 76.44 hours in total).

[154] Cormack & Mojdeh, *supra* note 98, at 3.

[155] *Id.* at 3.

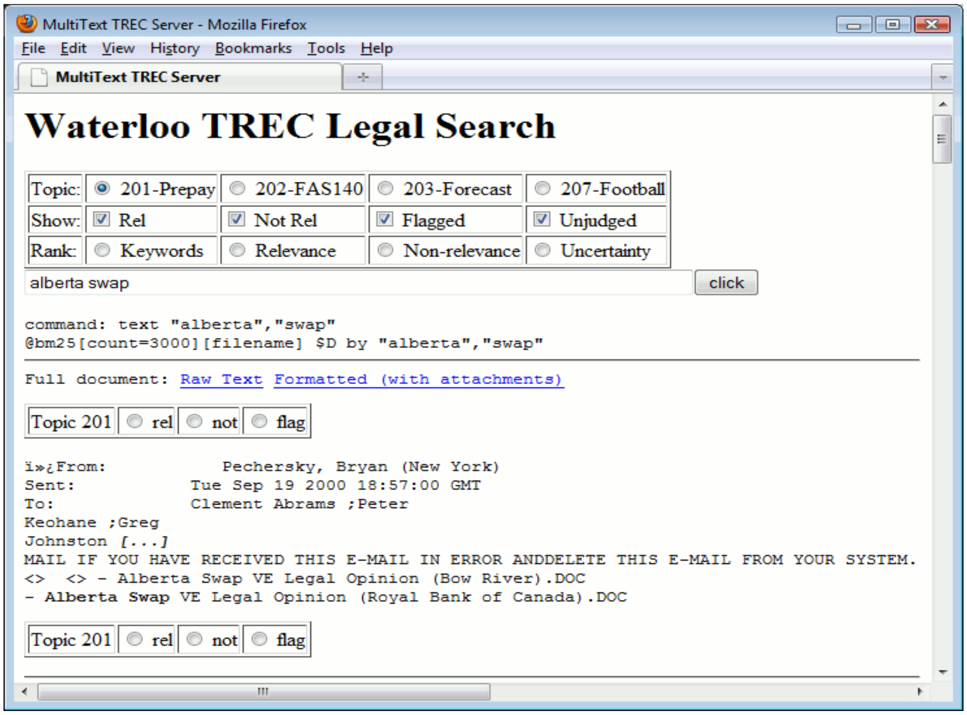
[156] *Id.*

[157] *Id*.

[158] *Id.*

[159] Cormack & Mojdeh, *supra* note 98, at 3.

[160] *Id.*


MultiText TREC Server - Mozilla Firefox
File Edit View History Bookmarks Tools Help
MultiText TREC Server
click

Waterloo TREC Legal Search

Topic:	● 201-Prepay	○ 202-FAS140	○ 203-Forecast	○ 207-Football
Show:	☑ Rel	☑ Not Rel	☑ Flagged	☑ Unjudged
Rank:	○ Keywords	○ Relevance	○ Non-relevance	○ Uncertainty

alberta swap

```
command: text "alberta","swap"
@bm25[count=3000][filename] $D by "alberta","swap"
```

Full document: Raw Text Formatted (with attachments)

Topic 201	○ rel	○ not	○ flag

```
ï»¿From:           Pechersky, Bryan (New York)
Sent:          Tue Sep 19 2000 18:57:00 GMT
To:            Clement Abrams ;Peter
Keohane ;Greg
Johnston [...]
MAIL IF YOU HAVE RECEIVED THIS E-MAIL IN ERROR ANDDELETE THIS E-MAIL FROM YOUR SYSTEM.
<>  <> - Alberta Swap VE Legal Opinion (Bow River).DOC
- Alberta Swap VE Legal Opinion (Royal Bank of Canada).DOC
```

Topic 201	○ rel	○ not	○ flag

Figure 1: Waterloo's interactive search and judging interface.[161]

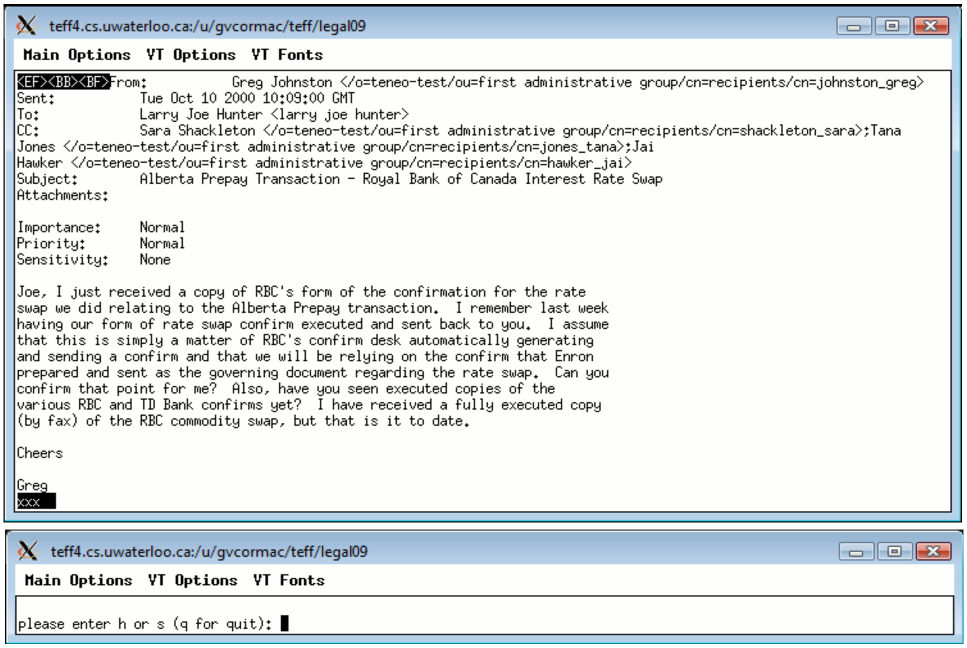
[42] The third and final phase estimated the density of relevant documents as a function of the score assigned by the active learning system, based on the assessments rendered during the active learning phase.[162] Waterloo used this estimate to gauge the tradeoff between recall and precision, and to determine the number of documents to produce so as to optimize F_1, as required by the task guidelines.[163]

[161] *Id.* at 3 & fig.2.

[162] *See id.* at 6.

[163] *Id.* at 3, 6; *see* Hedin et al., *supra* note 9, at 3.

[43] For Waterloo's TREC 2009 effort, the end result was that a human reviewed every document produced;[164] however, the number of documents reviewed was a small fraction of the entire document population (14,396 of the 836,165 documents were reviewed, on average, per topic).[165] Total review time for all phases was about 118 hours; 30 hours per topic, on average.[166]



```
<EF><BB><BF>From:           Greg Johnston </o=teneo-test/ou=first administrative group/cn=recipients/cn=johnston_greg>
Sent:           Tue Oct 10 2000 10:09:00 GMT
To:             Larry Joe Hunter <larry joe hunter>
CC:             Sara Shackleton </o=teneo-test/ou=first administrative group/cn=recipients/cn=shackleton_sara>;Tana
Jones </o=teneo-test/ou=first administrative group/cn=recipients/cn=jones_tana>;Jai
Hawker </o=teneo-test/ou=first administrative group/cn=recipients/cn=hawker_jai>
Subject:        Alberta Prepay Transaction - Royal Bank of Canada Interest Rate Swap
Attachments:

Importance:     Normal
Priority:       Normal
Sensitivity:    None

Joe, I just received a copy of RBC's form of the confirmation for the rate
swap we did relating to the Alberta Prepay transaction.  I remember last week
having our form of rate swap confirm executed and sent back to you.  I assume
that this is simply a matter of RBC's confirm desk automatically generating
and sending a confirm and that we will be relying on the confirm that Enron
prepared and sent as the governing document regarding the rate swap.  Can you
confirm that point for me?  Also, have you seen executed copies of the
various RBC and TD Bank confirms yet?  I have received a fully executed copy
(by fax) of the RBC commodity swap, but that is it to date.

Cheers

Greg
xxx
```

```
please enter h or s (q for quit): █
```

Figure 2: Waterloo's minimalist review interface.[167]

[164] *See* Cormack & Mojdeh *supra* note 98, at 6 ("the optimal strategy was to include *no* unassessed documents").

[165] *Id.*, at 6 tbl.2; E-mail from Bruce Hedin to Gordon V. Cormack, *supra* note 101 ("I have attached full list of the 836,165 document-level IDs").

[166] 118 hours is the sum of 76 hours for the interactive search and judging phase (*supra* para. 39) and 42 hours for the active learning phase (*supra* para. 41). Since Waterloo did four topics, the average effort per topic was 29.5 hours.

[167] Cormack & Mojdeh, *supra* note 98, at 4 fig.3.

IV. QUANTITATIVE ANALYSIS

[44] This Article's purpose 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 than exhaustive manual review, as measured by recall, precision, and F_1. To compare technology-assisted to manual review, the study required:

1. The results of one or more technology-assisted reviews. For this purpose, the Authors used the H5 review and the four Waterloo reviews conducted during the course of their participation in the TREC 2009 Legal Track Interactive Task.[168]

2. The results of manual reviews for the same topics and datasets as the technology-assisted reviews. For this purpose, the Authors used the manual reviews that TREC conducted on document samples for the purpose of evaluating the results that the participating teams submitted.[169]

3. A gold standard determination of relevance or nonrelevance. For this purpose, the Authors used the TREC final adjudicated assessments, for which the TA was the ultimate arbiter.[170]

[45] The Authors evaluated the results of the technology-assisted reviews and the manual reviews in exactly the same manner, using the

[168] The TREC results are available online, but use, dissemination and publication of the material is limited. Text REtrieval Conference (TREC), *Past Results*, NAT'L INST. STANDARDS & TECH., http://trec.nist.gov/results.html (last visited Apr. 11, 2011) ("Individuals may request access to the protected area containing the raw results by contacting the TREC Program Manager. Before receiving access, individuals will be asked to sign an agreement that acknowledges the limited uses for which the data can be used.").

[169] Text REtrieval Conference (TREC), *Relevance Judgments and Evaluation Tools for the Interactive Task*, NAT'L INST. STANDARDS & TECH., http://trec.nist.gov/data/legal/09/evalInt09.zip (last visited Apr. 11, 2011).

[170] *Id.*; *see* Hedin et al., *supra* note 9, at 2-3.

TREC methodology and the TREC gold standard.[171] To compare the effectiveness of the reviews, this Article reports, for each topic:

1. Recall, precision, and F_1 for both the technology-assisted and manual reviews.[172]

2. The *difference* in recall, the difference in precision, and the difference in F_1 between the technology-assisted and manual reviews.[173]

3. The *significance of the difference* for each measure, expressed as P.[174] 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.[175] The Authors used 100 bootstrap samples of paired differences to estimate the standard error of measurement, assuming a two-tailed normal distribution, to compute P.[176]

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 is easily

[171] *See* Hedin et al., *supra* note 9, at 2-5.

[172] *See id.* at 3 (reporting recall, precision, and F_1 for TREC participants); *infra* Table 7 (reporting recall, precision, and F_1 for the TREC manual reviews).

[173] *See infra* Table 7. A positive difference in some measure indicates that the technology-assisted review is superior in that measure, while a negative difference indicates that it is inferior.

[174] BÜTTCHER ET AL., *supra* note 19, at 426.

[175] *See id.*

[176] *See id.* at 412-31. "The *bootstrap* . . . is a method for simulating an empirical distribution modeling *f* (S) by sampling the sample *s*."). *Id.* at 424.

computed from the table. For example, for Topic 201, the difference in recall between Waterloo and TREC is 77.8% – 75.6% = +2.2%.

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 approaches (H5 and Waterloo) compared to exhaustive manual reviews (TREC). Results marked (*) are superior and overwhelmingly significant ($P < 0.001$). Results marked (†) are superior but not statistically significant ($P > 0.1$).[177]

[46] 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 ($P > 0.1$). As Table 7 illustrates, all of the measured differences are either overwhelmingly significant or not statistically significant.[178]

V. QUALITATIVE ANALYSIS

[47] The quantitative results show that the recall of the manual reviews varies from about 25% (Topic 203) to about 80% (Topic 202). That is, human assessors missed between 20% and 75% of all relevant documents.[179] Is this shortfall the result of clerical error, a

[177] For the information contained in this table, see *Past Results*, *supra* note 168; *Relevance Judgments and Evaluation Tools for the Interactive Task, supra* note 169. For details on the calculation and meaning of *P*, see s*upra* para. 43.

[178] *Supra* Table 7.

[179] *See supra* Table 7.

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 Rule 37(a)(4), such a production could constitute a failure to respond.[180]

[48] To address this question, the Authors examined the documents that the TREC assessors coded as nonresponsive to Topics 204 and 207, but H5 or Waterloo coded as responsive, and the TA adjudicated as responsive. Recall from Table 5 that Topic 204 concerned shredding and destruction of documents, while Topic 207 concerned football and gambling. The Authors chose these topics because they were more likely to be easily accessible to the reader, as opposed to other topics, which were more technical in nature. In addition, lawyers employed by professional review companies assessed these two topics using accepted practices for manual review.[181]

[49] For Topic 204, 160 of the assessed documents were coded as nonresponsive by the manual reviewers and responsive by H5 and the TA;[182] Topic 207, 51 documents met these same criteria except that Waterloo and the TA made the responsiveness determinations.[183] From these numbers, the Authors extrapolated that the manual reviewers would

[180] *See* FED. R. CIV. P. 37(a)(4).

[181] *See* Hedin et al., *supra* note 9, at 8 ("The review of the samples for three of the seven Interactive topics (203, 204, and 207) was carried out by two firms that include professional document-review services among their offerings.").

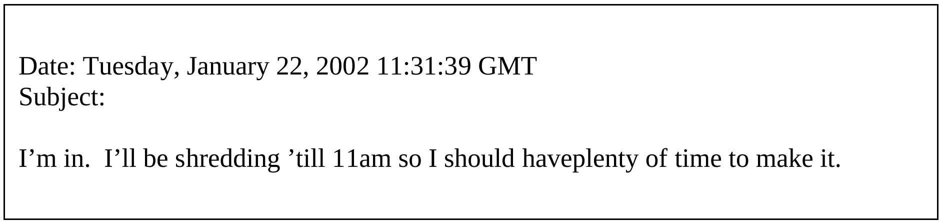
[182] The Authors identified these documents by comparing the submitted results, *see* Past *Results*, *supra* note 168 (file input.H52009.gz), the first-pass assessments, *see Relevance Judgments and Evaluation Tools for the Interactive Task*, *supra* note 169 (file qrels_doc_pre_all.txt), and the final adjudicated results, *see id.* (file qrels_doc_post_all.txt).

[183] The Authors identified these documents by comparing the submitted results, *see Past Results*, *supra* note 168 (file input.watlint.gz), the first-pass assessments, *see Relevance Judgments and Evaluation Tools for the Interactive Task*, *supra* note 169 (file qrels_doc_pre_all.txt), and the final adjudicated results, *see id.* (file qrels_doc_post_all.txt).

have missed 1,918 and 1,273 responsive documents (for Topics 204 and 207, respectively), had they reviewed the entire document collection.

[50] For each of these documents, the Authors used their judgment to assess whether the document had been miscoded due to:

- *Inarguable error*: Under any reasonable interpretation of relevance, the reviewer should have coded the document as responsive, but did 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.[184] 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.



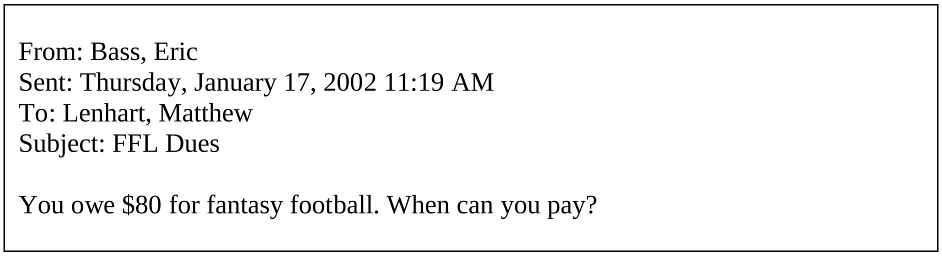
Date: Tuesday, January 22, 2002 11:31:39 GMT
Subject:

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

Figure 3: Topic 204 Inarguable error. A professional reviewer coded this document as nonresponsive, although it clearly pertains to document shredding, as specified in the production request.[185]

[184] *Cf.* Jeremy M. Wolfe et al., *Low Target Prevalence Is a Stubborn Source of Errors in Visual Search Tasks*, 136 J. EXPERIMENTAL PSYCH. 623, 623-24 (2007) (showing that in visual search tasks, humans have much higher error rates when the prevalence of target items is low).

[185] *See supra* Table 5. Figure 3 is an excerpt from document 0.7.47.1449689 in the TREC 2009 dataset, *supra* note 101.



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. A professional reviewer coded this document as nonresponsive, although it clearly pertains to fantasy football, as specified in the production request.[186]

- *Interpretive error*: Under some reasonable interpretation of relevance – but not the TA's interpretation as provided in the topic guidelines – an assessor might consider the document as nonresponsive. For example, a reviewer might have construed an automated message stating, "your mailbox is nearly full; please delete unwanted messages" (see Figure 5) as nonresponsive to Topic 204, although the TA defined it as responsive. Similarly, an assessor might have construed a message concerning children's football (see Figure 6) as nonresponsive to Topic 207, although the TA defined it as responsive.

[186] *See supra* Table 5. Figure 4 is an excerpt from document 0.7.47.320807 from the TREC 2009 dataset, *supra* note 101.

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. A professional reviewer coded this automated message as nonresponsive, although the TA construed such messages to be responsive to Topic 204.[187]

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 reviewer may have construed a children's league football game to be outside of the scope of "gambling on football." The TA deemed otherwise.[188]

- *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 (e.g., Figures 7 and 8).

[187] *See supra* Table 5. Figure 5 is an excerpt from document 0.7.47.1048852 in the TREC 2009 dataset, *supra* note 101.

[188] *See supra* Table 5. Figure 6 is an excerpt from document 0.7.47.668065 in the TREC 2009 dataset, *supra* note 101.

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. Applying the TA's conception of relevance, reasonable, informed assessors might disagree as to its responsiveness.[189]

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 reviewers might disagree about whether or not it is responsive according to the TA's conception of relevance.[190]

[51] When rendering assessments for the qualitative analysis, the Authors considered the mock complaint,[191] the topics,[192] and the topic-specific assessment guidelines memorializing the TA's conception of relevance, which were given to the human reviewers for reference

[189] *See supra* Table 5. Figure 7 is an excerpt from document 0.7.47.1304583 in the TREC 2009 dataset, *supra* note 101.

[190] *See supra* Table 5. Figure 8 shows an excerpt from document 0.7.6.179483 in the TREC 2009 dataset, *supra* note 101.

[191] *See generally Complaint, Grumby v. Volteron Corp.*, *supra* note 97.

[192] *Id.* at 14; Hedin et al, *supra* note 9, at 5-6.

purposes.[193] Table 8 summarizes the findings: The vast majority of missed documents are attributable either to inarguable error or to misinterpretation of the definition of relevance (interpretive error). Remarkably, the findings identify only 4% of all errors as arguable.

	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 that human reviewers missed, 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 missed documents, in the opinion of the Authors, have debatable responsiveness, according to the topic-specific guidelines.[194]

VI. RESULTS AND DISCUSSION

[52] 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 reviews. The technology-assisted reviews require, on average, human review of only 1.9% of the documents, a fifty-fold savings over exhaustive manual review. For F_1 and precision, the measured difference is overwhelmingly statistically significant ($P < 0.001$);[195] for recall the measured difference is not significant ($P > 0.1$).[196] These measurements provide strong evidence that the technology-assisted

[193] Text REtrieval Conference (TREC), *TREC-2009 Legal Track – Interactive Task, Topic-Specific Guidelines – Topic 204*, U. WATERLOO, http://plg1.cs.uwaterloo.ca/trec-assess/TopicGuidelines_204.pdf (last updated Oct. 22, 2009); Text REtrieval Conference (TREC), *TREC-2009 Legal Track – Interactive Task, Topic-Specific Guidelines – Topic 207*, U. WATERLOO, http://plg1.cs.uwaterloo.ca/trec-assess/TopicGuidelines_207_.pdf (last updated Oct. 22, 2009).

[194] *See* sources cited *supra* note 193.

[195] *See supra* Tables 6, 7.

[196] *Id*.

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.

[53] It should be noted that the objective of TREC participants was to maximize F_1, not recall or precision, per se.[197] It happens that they achieved, on average, higher precision.[198] 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.

[54] Table 7 shows that, for four of the five topics, the technology-assisted processes achieve substantially higher F_1 scores, largely due to their high precision. Nonetheless, for a majority of the topics, the technology-assisted processes achieve higher recall as well; for two topics, substantially higher.[199] For Topic 207, there is no meaningful difference in effectiveness between the technology-assisted and manual reviews, for any of the three measures. *There is not one single measure for which manual review is significantly better than technology-assisted review.*

[55] For three of the five topics (Topics 201, 202, and 207) the results show no significant difference in recall between the technology-assisted and manual reviews. This result is perhaps not surprising, since the recall scores are all on the order of 70% – the best that might be reasonably achieved, given the level of agreement among human assessors. As such, the 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.

[197] *See* Hedin et al., *supra* note 9, at 15.

[198] *See supra* Tables 6, 7.

[199] *See supra* Table 7.

VII. Limitations

[56] The 2009 TREC effort used a mock complaint and production requests composed by lawyers to be as realistic as possible.[200] Furthermore, the role of the TA was intended to simulate that of a senior attorney overseeing a real document review.[201] Finally, the dataset consisted of real e-mail messages captured within the context of an actual investigation.[202] These components of the study are perhaps as realistic as might reasonably be achieved outside of an actual legal setting.[203] 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.[204] Both participants and TAs may have had prior knowledge of both the story and dataset, affecting 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,[205] books,[206]

[200] Hedin et al., *supra* note 9, at 2.

[201] *See id.*; *see also* Oard et al., *supra* note 9, at 20.

[202] *See* Hedin et al., *supra* note 9, at 4.

[203] *See id.*

[204] *See, e.g.*, John Markoff, Armies of Expensive Lawyers Replaced by Cheaper Software, N.Y. TIMES, Mar. 5, 2011, A1, available at http://www.nytimes.com/2011/03/05/science/05legal.html; *see also* E-mail from Jonathan Nystrom to Maura R. Grossman (Apr. 5, 2011 19:12 EDT) (on file with authors) (confirming use of the Enron data set for product demonstrations); E-mail from Jim Renehan to Maura R. Grossman (Apr. 5, 2011 20:06 EDT) (on file with authors) (confirming use of the Enron data set for product demonstrations); E-mail from Lisa Schofield to Maura R. Grossman (Apr. 5, 2011 18:27 EDT) (on file with authors) (confirming use of the Enron data set for product demonstrations); E-mail from Edward Stroz to Maura R. Grossman (Apr. 5, 2011 18:32 EDT) (on file with authors) (confirming use of the Enron data set for product demonstrations).

[205] *See, e.g.*, John C. Coffee Jr., *What Caused Enron?: A Capsule Social and Economic History of the 1990's*, 89 CORNELL L. REV. 269 (2004); Paul M. Healy & Krishna G. Palepu, *The Fall of Enron*, 17 J. ECON. PERSP. 3 (2003).

and even a popular movie.[207] It is unclear what effect, if any, these factors may have had on the results.

[57] 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 required the production of documents related to "transactions that the Company characterized as compliant with FAS 140."[208] Participating teams were required to undertake research to identify the relevant transactions, as well as the names of the parties, counterparties, and entities involved.[209] Manual reviewers, on the other hand, were given detailed guidelines specifying these elements.[210]

[58] Moreover, TREC conducted manual review on a stratified sample containing a higher proportion of relevant documents than the collection as a whole,[211] and used statistical inference to evaluate the result of reviewing every document in the collection.[212] Beyond the statistical uncertainty, there also 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

[206] *See, e.g.*, LOREN FOX, ENRON: THE RISE AND FALL (2002); BETHANY MCLEAN AND PETER ELKIND, THE SMARTEST GUYS IN THE ROOM: THE AMAZING RISE AND SCANDALOUS FALL OF ENRON (2003).

[207] ENRON: THE SMARTEST GUYS IN THE ROOM (Magnolia Pictures 2005).

[208] Hedin et al., *supra* note 9, at 5.

[209] *See id*. at 8.

[210] *See id*. at 3.

[211] *See id*. at 12, tbl.3.

[212] *See generally id*.

sampling effect, combined with the greater guidance provided to the human reviewers, may have resulted in an overestimate of the effectiveness of manual review, and thus understated the results of the study.

[59] Of note is the fact that the appeals process involved reconsideration – and potential reversal – *only* of manual coding decisions that one or more participating teams appealed, presumably because their results disagreed with the manual reviewers' decisions.[213] The appeals process depended on participants exercising due diligence in identifying the assessments with which they disagreed.[214] And while it appears that H5 and Waterloo exercised such diligence, it became apparent to the Authors during the course of their analysis that a few assessor errors were overlooked.[215] These erroneous assessments were deemed correct under the gold standard, with the net effect of overstating the effectiveness of manual reviews, while understating the effectiveness of technology-assisted review.[216] It is also likely that the manual review and technology-assisted processes incorrectly coded some documents that were not appealed.[217] The impact of the resulting errors on the gold standard would be to overstate both recall and precision for manual review, as well as for technology-assisted review, with no net advantage to either.

[213] *See* Hedin et al., *supra* note 9 at 3, 13-14. There is no benefit, and therefore no incentive, for participating teams to appeal coding decisions with which they agree.

[214] *See id.* If participating teams do not appeal the manual reviewers' incorrect decisions, those incorrect decisions will be incorporated into the gold standard, compromising its accuracy and usefulness.

[215] Hedin et al., *supra* note 9 at 14, tbl.4 (showing that for every topic, H5 and Waterloo appealed the majority of disagreements between their results and the manual assessments).

[216] *See supra* note 214. If the manual review is incorrect, and the technology-assisted review is correct, the results will overstate the effectiveness of manual review at the expense of technology-assisted review.

[217] Given that neither the manual reviewers nor the technology-assisted processes are infallible, it stands to reason that they may occasionally agree on coding decisions that are incorrect.

[60] In designing this study, the Authors considered only the results of two of the eleven teams participating in TREC 2009, because they were considered most likely to demonstrate that technology-assisted review can improve upon exhaustive manual review. The study considered all submissions by these two teams, which happened to be the most effective submissions for five of the seven topics. The study did not consider Topics 205 and 206, because neither H5 nor Waterloo submitted results for them. Furthermore, due to a dearth of appeals, there was no reliable gold standard for Topic 206.[218] The Authors were aware before conducting their analysis that the H5 and Waterloo submissions were the most effective for their respective topics. To show that the results are significant in spite of this prior knowledge, the Authors applied Bonferroni correction,[219] which multiplies P by 11, the number of participating teams. Even under Bonferroni correction, the results are overwhelmingly significant.

VIII. CONCLUSION

[61] 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.

[218] Hedin et al., *supra* note 9, at 17-18 ("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 . . . .").

[219] *See* BÜTTCHER ET AL., *supra* note 19, at 428.

Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review

Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review

Herbert L. Roitblat
Electronic Discovery Institute, OrcaTec LLC, PO Box 613, Ojai, CA 93024. E-mail: herb@orcatec.com

Anne Kershaw
Electronic Discovery Institute, A. Kershaw, P.C. Attorneys & Consultants, 303 South Broadway, Suite 430, Tarrytown, NY 10591. E-mail: anne.kershaw@akershaw.com

Patrick Oot
Electronic Discovery Institute, Verizon, 1320 North Courthouse Road, Arlington, VA 22201. E-mail: patrick.oot@verizon.com

In litigation in the US, the parties are obligated to produce to one another, when requested, those documents that are potentially relevant to issues and facts of the litigation (called "discovery"). As the volume of electronic documents continues to grow, the expense of dealing with this obligation threatens to surpass the amounts at issue and the time to identify these relevant documents can delay a case for months or years. The same holds true for government investigations and third-parties served with subpoenas. As a result, litigants are looking for ways to reduce the time and expense of discovery. One approach is to supplant or reduce the traditional means of having people, usually attorneys, read each document, with automated procedures that use information retrieval and machine categorization to identify the relevant documents. This study compared an original categorization, obtained as part of a response to a Department of Justice Request and produced by having one or more of 225 attorneys review each document with automated categorization systems provided by two legal service providers. The goal was to determine whether the automated systems could categorize documents at least as well as human reviewers could, thereby saving time and expense. The results support the idea that machine categorization is no less accurate at identifying relevant/responsive documents than employing a team of reviewers. Based on these results, it would appear that using machine categorization can be a reasonable substitute for human review.

Received June 5, 2009; revised July 29, 2009; accepted August 27, 2009

© 2009 ASIS&T • Published online 8 October 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.21233

Introduction

In litigation, particularly civil litigation in the US Federal Courts, the parties are required, when requested, to produce documents that are potentially relevant to the issues and facts of the matter. This is a part of the process called "discovery." When it involves electronic documents, or more formally, "electronically stored information (ESI)," it is called eDiscovery. The potentially relevant documents are said to be responsive.

The volume of electronically stored information that must be considered for relevance continues to grow and continues to present a challenge to the parties. The cost of eDiscovery can easily be in the millions of dollars. According to some commentators, these costs threaten to skew the justice system as the costs can easily exceed the amount at risk (Bace, 2007). Discovery is a major source of costs in litigation, sometimes accounting for as much as 25% of the total cost (e.g., Gruner, 2008). Overwhelmingly, the biggest single cost in eDiscovery is for attorney review time—the time spent considering whether each document is responsive (relevant) or not. Traditionally, each document or email was reviewed by an attorney who decided whether it was responsive or not. As the volume of material that needs to be considered continues to grow, it is becoming increasingly untenable to pursue that strategy.

Attorneys and their clients are looking for ways to minimize the cost of eDiscovery (Paul & Baron, 2007). One approach that holds promise for reducing costs while delivering appropriate results is the use of information retrieval tools.

Over the last several years, attorneys have come to rely increasingly on search tools, for example, Boolean queries, to limit the scope of what must be reviewed. The details of these queries may be negotiated between the parties. Here is an example of one such query in the case of U.S. v Philip Morris:

> (((master settlement agreement OR msa) AND NOT
> (medical savings account OR metropolitan standard area))
> OR s. 1415 OR
> (ets AND NOT educational testing service) OR
> (liggett AND NOT sharon a. liggett) OR atco OR lorillard
> OR
> (pmi AND NOT presidential management intern) OR pm usa
> OR rjr OR
> (b&w AND NOT photo* OR phillip morris OR batco OR ftc
> test method OR star scientific OR vector group OR joe
> camel OR
> (marlboro AND NOT upper marlboro)) AND NOT
> (tobacco* OR cigarette* OR smoking OR tar OR nicotine OR
> smokeless OR synar amendment OR philip morris OR r.j.
> reynolds OR
> ("brown and williamson") OR
> ("brown & williamson") OR bat industries OR liggett group)
> (Baron, 2008).

The information retrieval requirements of attorneys conducting eDiscovery are somewhat different from those in many information retrieval tasks. Document sets in eDiscovery tend to be very large with a large proportion of emails and a large number of requests that need to be translated into queries. The Philip Morris case, for example, involved over 1,726 requests from the tobacco companies and more than 32 million Clinton-era records that needed to be evaluated.

Information retrieval studies involving the World Wide Web, of course, have an even greater population of potentially relevant documents, but in those systems the user is usually interested in only a very tiny proportion of them, for example, between 1 and 50 documents out of billions. Getting the desired information within the first 10–50 results is generally the challenge in these studies.

Web searches are generally fairly specific, for example, "What are the best sites to visit in Paris?" In contrast, the information need in eDiscovery is generally much broader and more vague. Discovery requests include statements like "All documents constituting or reflecting discussions about unfair or discriminatory allocations of [Brand X] products or the fear of such unfair or discriminatory allocations." These requests will not typically be satisfied by one or a few documents.

Recall, the proportion of responsive documents actually retrieved, is arguably a more important measure of the success of information retrieval for the lawyers than is precision, the proportion of retrieved documents that are responsive. High precision will save the client money, because fewer documents will need to be reviewed. On the other hand, obviously low recall can lead to court sanctions, including an "adverse inference" instruction, where a jury is instructed that they may construe that the missing information was contrary to the interests of the party that failed to produce it. Obviously, low precision can also lead to accusations that the producing party is doing an inadequate job identifying responsive documents, but these sanctions are usually much less onerous than those for failing to produce.

This study is an investigation of methods that may be useful to reduce the expense and time needed to conduct electronic discovery. In addition to search techniques, these methods can include machine learning and other data mining techniques. In the present study, the categorization tools provided by two companies who are active eDiscovery service providers were used to categorize responsive documents. These providers' systems were taken to be representative of a broad range of similar systems that are available to litigators. The performance of these two systems was compared to the performance of a more traditional methodology—having attorneys read and categorize each document in the context of a substantial eDiscovery project.

Background: Related Work

Blair and Maron (1985) conducted one of the early studies on using computers to identify potentially responsive documents. They analyzed the search performance of attorneys working with experienced search professionals to find documents relevant to a case in which a computerized San Francisco Bay Area Rapid Transit (BART) train failed to stop at the end of the line. The case involved a collection which, at the time, seemed rather large, consisting of about 40,000 documents. Current cases often involve one to two orders of magnitude more documents.

Blair and Maron found that the attorney teams were relatively ineffective at using the search system to find responsive documents. Although they thought that their searches had retrieved 75% or more of the responsive documents, they had, in fact, found about 20% of them.

One reason for this difficulty is the variety of language used by the parties in the case. The parties on the BART side referred to "the unfortunate incident," but parties on the victim's side called it an "accident" or a "disaster." Some documents referred to the "event," "incident," "situation," "problem," or "difficulty." Proper names were sometimes left out. The limitation in this study was not the ability of the computer to find documents that met the attorneys' search criteria, but the inability to anticipate all of the possible ways that people could refer to the issues in the case.

Blair and Maron concluded that "It is impossibly difficult for users to predict the exact words, word combinations, and phrases that are used by all (or most) relevant documents and only (or primarily) by those documents" (p. 295). They advocated for the use of manually applied index terms, meaning that someone would have to read the documents, determine what they were about, and categorize them.

TREC (Text Retrieval Conference) is a multitrack project sponsored by the National Institute for Standards and Technology and others to conduct comparative research on text retrieval technologies. Since 2006 (Baron, Lewis, & Oard, 2007; Tomlinson, Oard, Baron, & Thompson, 2008, Oard, Hedin, Tomlinson, & Baron, 2009), TREC has included a legal track whose goal is to assess the ability of information retrieval technology to "meet the needs of the legal community for tools to help with retrieval of business records." In support of this goal, they seek to develop and apply collections and tasks that approximate the data, methods, and issues that real attorneys might use during civil litigation and to apply objective criteria by which to judge the success of various search methodologies. In 2008 (Oard et al., 2009), 15 research teams participated in at least one of the three types of task (ad hoc query, relevance feedback, and interactive search).

The searches were conducted against a collection (also used in 2006 and 2007) of tobacco-related documents released under the Tobacco Master Settlement Agreement (MSA) called the IIT Complex Document Information Processing Test Collection (CDIP) v. 1.0. The collection consists of 6,910,192 document records in the form of XML elements. Most of these documents were encoded from images using optical character recognition (OCR). Relying on OCR data for text presents its own challenges to these studies, because of the less than perfect accuracy of the process used to derive the text from the documents.

The performance of the various teams on each task was measured by having a pool of volunteer assessors evaluate a sample of documents for relevance. The assessors for the 2008 session were primarily second- and third-year law students, with a few recent law school graduates, experienced paralegals, and other litigation specialists. Each assessor was asked to evaluate a minimum of 500 documents. On average, an assessor managed about 21.5 documents per hour, so a block of 500 documents entailed a substantial level of effort from the volunteer assessors.

In the TREC ad hoc task, the highest recall achieved was 0.555 (i.e., 55.5% of the documents identified as relevant were retrieved; Run "wat7fuse"). The precision corresponding to that level of recall was 0.210, meaning that 21% of the retrieved documents were determined to be relevant.

The TREC interactive task allowed each team to interact with a topic authority and revise their queries based on this feedback. Each team was allowed 10 hours of access to the authority. The interactive task also allowed the teams to appeal reviewer decisions if they thought that the reviewers had made a mistake. Of the 13,339 documents that were assessed for the interactive task, 966 were appealed to the topic authority. This authority played the role, for example, of the senior litigator on the case, with the ultimate authority to overturn the decisions of the volunteer assessors. In about 80% of these appeals the topic authority supported the appeal and recategorized the document. In one case (Topic 103), the appeal allowed the team with the already highest recall rate to improve its recall by 47%, ending up with recall of 0.624 and precision of 0.810.

Some of the more interesting findings from the 2008 TREC legal track concern the levels of agreement seen between assessors. Some of the same topics were used in previous years of the TREC legal track, so it is possible to compare the judgments made during the current year with those made in previous years. For example, the level of agreement between assessors in the 2008 project and those from 2006 and 2007 were reported. Ten documents from each of the repeated topics that were previously judged to be relevant and 10 that were previously judged to be nonrelevant were assessed by the 2008 assessors. It turns out that "just 58% of previously judged relevant documents were judged relevant again this year." Conversely, "18% of previously judged non-relevant documents were judged relevant this year." Overall, the 2008 assessors agreed with the previous assessors 71.3% of the time.

Unfortunately, this is a fairly small sample, but it is consistent with other studies of inter-reviewer agreement. In 2006 the TREC coordinators gave a sample of 25 relevant and 25 nonrelevant documents from each topic to a second assessor and measured the agreement (http://cio.nist.gov/esd/emaildir/lists/ireval/msg00012.html, retrieved 23 July, 2009) between these two. Here they found about 76% agreement. Other studies outside of TREC Legal have found similar levels of (dis)agreement (e.g., Barnett, Godjevac, Renders, Privault, Schneider, & Wickstrom, 2009; Borko, 1964; Tonta, 1991; Voorhees, 1998).

Research Design: Methods

Research Questions

One solution to the problem of the exploding cost of eDiscovery is to use technology to reduce the effort required to identify responsive and privileged documents. Like the TREC legal track, the goal of the present research is to evaluate the ability of information retrieval technology to meet the needs of the legal community for tools to identify the responsive documents in a collection.

From a legal perspective, there is recognition that the processes used in discovery do not have to be absolutely perfect, but should be reasonable and not unduly burdensome (e.g., Rule 26(g) of the Federal Rules of Civil Procedure). The present study is intended to investigate whether the use of technology is reasonable in this sense.

The notion of "reasonable" is itself subject to interpretation. We have taken the approach that the current common practice of having trained reviewers examine each document does a reasonable job of identifying responsive documents, but at an often unreasonable cost. If information retrieval systems can be used to achieve the same level of performance as the current standard practice, then they too should be considered reasonable by this standard. 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.*

Participants

The participants in this study were the original review teams, two re-review teams, and two electronic discovery service providers. The original review was conducted by two teams of attorneys, one focused on review for privilege, and one focused on review for relevance. A total of 225 attorneys participated in this initial review. The original purpose of this review was to meet the requirements of a US Department of Justice investigation of the acquisition of MCI by Verizon. It was not initially designed as a research study, but Verizon has made the outcome of this review available in support of the present study. For more details, see the Dataset section, below.

The two re-review teams were employees of a service provider specializing in conducting legal reviews of this sort. Each team consisted of five reviewers who were experienced in the subject matter of this collection. The two teams of re-reviewers (Team A and Team B) both reviewed the same 5,000 documents in preparation for one of the processes of one of the two service providers. Hence, there is a caveat that the decisions made by the service provider are not completely independent of the decisions made by the re-review teams. This issue will be discussed further in the Discussion section.

The two service providers volunteered their time, facilities, and processes to analyze the data. The two companies, one based in California and the other in Texas, each independently analyzed the data without knowledge of the original decisions made or of the decisions made by the other provider. Their systems are designated System C and System D. The identity of the two systems, that is, which company's is System C and which is System D, was determined by a coin flip in order to conceal the identity of the system yielding specific data. We did not cast this task as a competition between the two systems and do not wish to draw distinctions between them. Rather, we see these two systems as representative of a general analytic approach to information retrieval in electronic discovery.

Task

The task of the original review was to determine whether each document was responsive to the request of the Justice Department. The reviewers also made decisions about the privilege status of the documents, but these judgments were not used in the present study.

The task of the two systems was to replicate the classification of documents into the two categories of responsive and nonresponsive.

Dataset

The documents used in the present study were collected in response to a "Second Request" concerning Verizon's acquisition of MCI. Certain corporate mergers and acquisitions are subject to premerger antitrust review by the U.S. Department of Justice and the U.S. Federal Trade Commission under the Hart-Scott-Rodino Antitrust Improvements Act of 1976. The government may issue requests for additional information for material relevant to the proposed acquisition. These requests are commonly called "Second Requests."

The documents were collected from 83 employees in 10 US states. Together they consisted of 1.3 terabytes of electronic files in the form of 2,319,346 documents. The collection consisted of about 1.5 million email messages, 300,000 loose files, and 600,000 scanned documents. After eliminating duplicates, 1,600,047 items were submitted for review. The attorneys spent about 4 months, 7 days a week, and 16 hours per day on the review at a total cost of $13,598,872.61 or about $8.50 per document. After review, a total of 176,440 items were produced to the Justice Department.

During the course of managing the Second Request, these documents were put in a database, containing 2,224,023 records. From this database, 1,813,612 documents could be usably extracted for this study.

Analysis

This study compares two computer-assisted categorization processes with the more traditional process of having people, usually lawyers, read and categorize each document. This study uses agreement to assess the level of reliability of the human and computer processes.

A sample of 5,000 documents from the original review was randomly chosen for a second review ~2 years after complying with the Second Request. Two teams, A and B, performed this re-review without explicit knowledge of the original classifications. Both teams reviewed the complete sample.

The goal of the analysis is to compare the level of agreement between the two re-review teams and the two computer systems, on the one hand, with the original review on the other. The level of agreement between the two re-review teams and the original review is an indication of the level of agreement one could expect from the traditional process. To the extent that the computer systems show similar levels of agreement, they could be thought of as a reasonable substitute for the traditional process.

Results

The original review identified 9.46% ($SE = 0.0002$) of the documents in the entire collection and 9.8% ($SE = 0.0042$) of the documents in the sample of 5,000 as responsive. The difference between these two proportions was not significant ($t = 0.719$, $p > 0.05$), supporting the notion that the sample was representative of the complete population.

The contingency tables showing how the different systems and reviewers classified documents are shown in Table 1. Each row represents the contingent relation shown in the first column. RR indicates the number of documents that were classified as responsive by both groups or systems. RN indicates the number of documents that were classified as responsive by the first group and as nonresponsive by the second group. NR indicates the documents that were classified as nonresponsive by the first group or system and as

DOI: 10.1002/asi

TABLE 1. The contingent frequencies between the designated sets of judgments.

	Contingent relation			
	RR	RN	NR	NN
1. Original vs. Team A	238	250	971	3,541
2. Original vs. Team B	263	225	1,175	3,337
3. Team A vs. Team B	580	629	858	2,933
4. Original vs. Teams A & B Nonadjudicated	349	139	1,718	2,794
5. Original vs. Teams A & B Adjudicated	216	272	739	3,773
6. Original vs. System C	78,617	92,908	211,403	1,430,684
7. Original vs. System C	90,416	81,109	216,359	1,425,728

Note. RR = Responsive/Responsive, RN = Responsive Nonresponsive, NR = Nonresponsive/Responsive, NN = Nonresponsive/Nonresponsive.

responsive by the second. NN indicates the documents that were classified as nonresponsive by both groups or systems.

Human Review

The contingency tables resulting from each of the two teams, compared with the original classifications, are shown in the first two rows of Table 1. Both contingency tables were significantly different from chance (independence) (Team A: $\chi^2 = 178.37$, Team B: $\chi^2 = 166.73$, both $p < 0.01$).

Row 3 of Table 1 shows the contingency table comparing Team B's classifications with those from Team A. The decisions made by the two teams were strongly related ($\chi^2 = 287.31$, $p < 0.01$).

The 1,487 documents on which Teams A and B disagreed 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. Row 4 of Table 1 contains the contingency table comparing the nonadjudicated decisions to the original classification and Row 5 contains the contingency table comparing the adjudicated decisions to the original classification. The adjudicated decisions, like those made independently by the two teams, were strongly related ($\chi^2 = 203.07$, $p < 0.01$) to the original review.

Team A identified 24.2% (SE = 0.006) and Team B identified 28.76% (SE = 0.006) of the sample as responsive. The difference between these two proportions was significant ($t = 5.20$, $p < 0.01$). After adjudication, the combined teams identified 955 or 19.1% (SE = 0.006) as responsive. Adjudication, in other words, reduced the overall number of documents that the new reviewers designated as responsive. Of the 1,487 documents on which Team A and Team B disagreed, the senior litigator chose Team A's classification on 796 documents, Team B's classification on 691 documents.

Team A agreed with the original review on 75.58% (SE = 0.006) of the documents and Team B agreed with the original review on 72.00% (SE = 0.006), both before adjudication. Team A agreed with Team B on 70.26% (SE = 0.006) of the documents. The adjudicated review agreed with the original classification on 79.8% (SE = 0.006) of the documents. Team A agreed with the original significantly more often ($t = 4.07$, $p < 0.01$) than did Team B. Because the adjudicated results included most of the decisions from Team A and Team B, it is not clear how to assess the difference in agreement between the adjudicated and nonadjudicated reviews—they are not independent.

Of the 488 documents in the sample identified as responsive by the original review team, Team A identified 238 or 48.78% (SE = 0.023) as responsive. Team B identified 263 or 53.89% (SE = 0.023) as responsive. Together, teams A and B identified as responsive 349 or 71.52% (SE = 0.02) of the documents classified as responsive by the original review. Conversely, of the 2067 documents identified as responsive by either Team A or Team B, the original review identified 349 or 16.88% (SE = 0.008) as responsive.

Of the 4,512 documents in the sample that were designated nonresponsive during the original review, Team A identified 971 or 21.52% (SE = 0.006) as responsive and Team B recognized 1,175 or 26.04% (SE = 0.007) as responsive. Together, Teams A and B recognized 1,718 or 38.07% (SE = 0.007) of these as responsive (before adjudication, i.e., if either team called it responsive, a document was counted for this purpose as responsive). After adjudication, the two teams combined recognized 739 or 16.38% (SE = 0.006) of the original review's nonresponsive documents as responsive.

Computer-Assisted Review and Comparison

In addition to the two review teams reexamining a sample of documents from the original review, two commercial electronic discovery systems were also used to classify documents as responsive vs. nonresponsive. 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 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.

The contingency table resulting from each of the two systems is shown in Rows 6 and 7 of Table 1.

DOI: 10.1002/asi

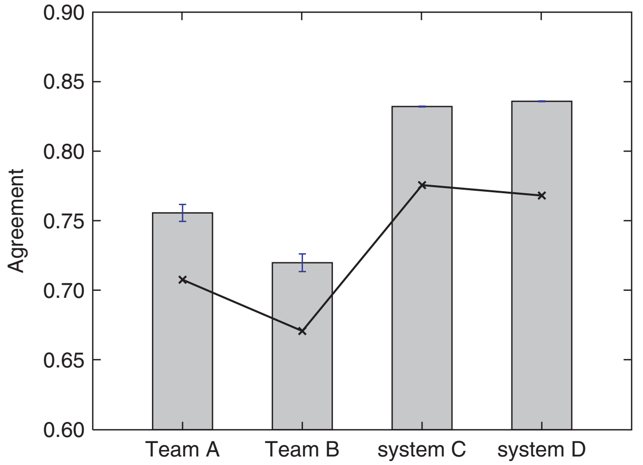


FIG. 1. The level of agreement with the original review and chance levels to be expected from the marginals for the two human teams and the two computer systems (the four reassessments). Error bars show standard error.

System C classified 15.99% (SE = 0.0003) of the documents and System D classified 16.92% (SE = 0.0003) of the documents as responsive, which were both higher than the proportion identified as responsive by the original team ($t = 187.6$, $p < 0.01$ and $t = 211.2$, $p < 0.01$, respectively). System C agreed with the original classification on 83.2% (SE = 0.00028) and System D agreed with the original classification on 83.6% (SE = 0.00028) of the documents.

Of the 171,525 documents identified as responsive by the original review team, System C identified 78,617 or 45.8% (SE = 0.001) as responsive. System D identified 90,416 or 52.7%% (SE = 0.001) as responsive. Together, Systems C and D identified as responsive (i.e., either C or D responsive), 123,750 or 72.1% (SE = 0.001) of the documents classified as responsive by the original review. Conversely, of the 493,004 documents identified as responsive by either System C or System D, the original review identified 123,750 or 25.1% (SE = 0.001) as responsive.

The percentage agreements between each of the two teams and each of the two systems and the original review are shown in Figure 1. The percentage agreements for each of the assessments shown in Figure 1 was significantly different from each other's assessment (A vs. B: $t = 4.07$, A vs. C: $t = 12.56$, A vs. D: 136.7, B vs. C: 17.65, B vs. D: 130.8, C vs. D: 2139.2, all $p < 0.01$). In addition, each assessment was significantly different from chance ($\chi^2 = 178.37$, 166.73, 125588.00, 172739.91, for A, B, C, and D, respectively, all $p < 0.01$).

Figure 2 breaks down overall agreement into positive agreement and negative agreement, proportions of specific agreement (Spitzer & Fleiss, 1974). When the base rates of the different categories are widely different, simple agreement is subject to chance-related bias. Positive and negative agreement remove that bias and allow one to look at each of these categories separately. Chance should affect only the more frequent category, in this case, the nonresponsive documents.

On positive agreement, assessments A and B did not differ significantly ($t = 0.702$, $p > 0.05$), but each of the other assessments differed from one another (t: A vs. C: 8.02, A vs. D: 50.10, B vs. C: 9.12, B vs. D: 50.79, C vs. D: 600.22, all $p < 0.01$). A similar pattern was seen for negative agreement. Assessments A and B did not differ significantly from one another ($p > 0.05$), but the comparisons did show significant differences in the degree to which they agreed with original review ($p < 0.01$) (t: A vs. B: 1.44, A vs. C: 3.89, A vs. D: 68.77, B vs. C: 6.00, B vs. D: 69.49, and C vs. D 905.68).

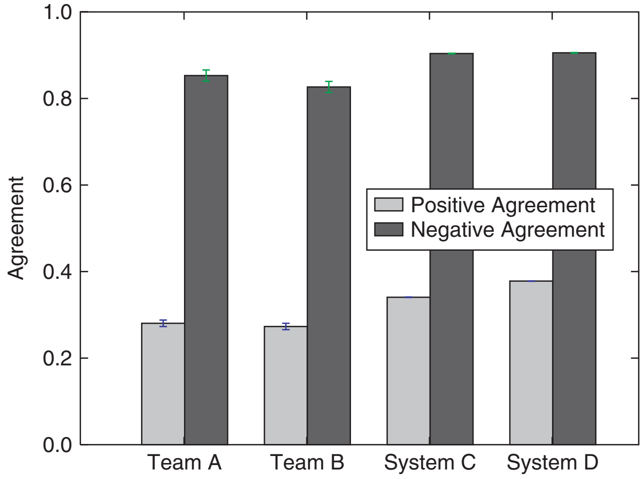


FIG. 2. Positive agreement (2*RR/(2*RR + RN + NR)) and negative agreement (2*NN/(2*NN + NR + RN)) for agreement between the original review and the four reassessments. NN, NR, etc. refer to the columns of Table 1. Error bars are standard error.

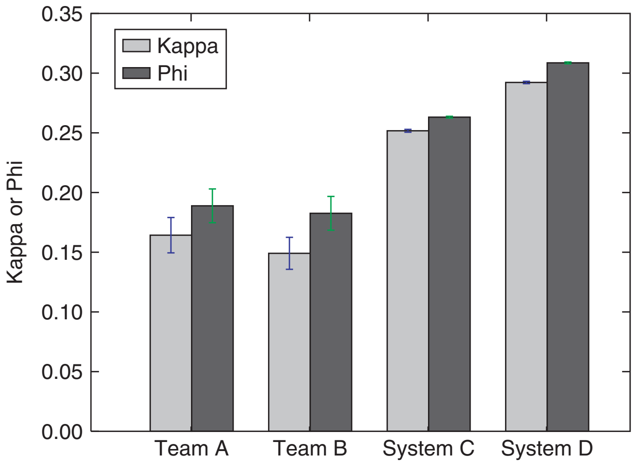


FIG. 3. Kappa and Phi for the agreement between the original review and the four reassessments. Kappa and Phi are "chance adjusted" measures of association or agreement. Error bars are standard error.

Another approach to characterizing the relationship between the latter assessments and the earlier reviews is to use "chance-corrected" measures of agreement. Figure 3 shows Cohen's kappa and phi, two measures that take into account the extent to which we might expect the assessments to agree based on chance. Cohen's kappa essentially subtracts out the level of agreement that one would expect by chance. Kappa is 1.0 if the two raters agree perfectly and is 0 if they agree exactly as often as expected by chance. Kappa less than 0 can be obtained if the raters agree less often than is expected by chance. Phi is derived from chi-squared and measures

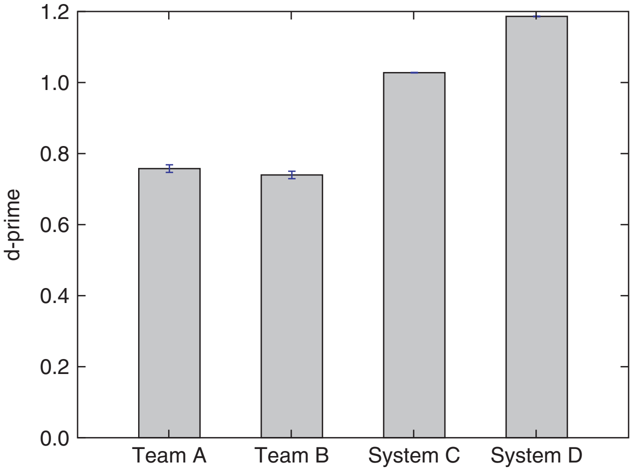


FIG. 4. The signal detection measure d′ comparing each of the re-reviews against the original review.

the deviation from the chance expectation. It has the value 0 only when there is complete independence between the two assessments. The pattern of results for both of these measures is the same as for agreement and for positive and negative agreement.

As with positive and negative agreement, Teams A and B did not differ significantly for either kappa ($t = 0.76$, $p > 0.05$) or phi ($t = 0.31$, $p > 0.05$). The other assessments did differ significantly from one another on kappa ($p < 0.01$) (t: A vs. C: 5.89, A vs. D: 8.62, B vs. C: 7.63, B vs. D: 10.65, C vs. D: 25.91) and on phi (t: A vs. C: 5.24, A vs. D: 8.45, B vs. C: 5.69, B vs. D: 8.90, C vs. D: 43.30). In addition to the data shown in Figure 3, we can also compute the corresponding measures comparing the decisions made by Team A with those made by Team B (kappa: 0.238, phi: 0.240).

The difference between the proportions identified as responsive by the original review and the re-reviews may indicate a difference in bias. Bias simply refers to an overall tendency to select one category over another, independent of the information in the documents. For example, one attorney might believe that it is more important to avoid missing a responsive document than another attorney does and so be more willing to classify documents as responsive. Recall increases and precision decreases when an assessor increases their willingness to call a document responsive; thus, these measures make it difficult to separate the discriminability of the classes from the bias. Signal detection theory (van Rijsbergen, 1979; Swets, 1969), on the other hand, offers a measure, d′, that is independent of bias. The more a system (or person) can separate two classes, the higher its d′ score will be. The value of d′ ranges from 0 when the responsive and nonresponsive documents are completely indistinguishable by the system to positive infinity when there is no overlap between the two.

With large numbers of trials (in our case documents), the binomial distribution is closely approximated by the normal distribution, so the use of the measure d′ is justified. Figure 4 shows the sensitivity measure, d′ for each of the four re-reviews.

TABLE 2. Standard information retrieval measures.

	Precision	Recall	F_1
Human Team A	0.196857	0.487705	0.280495
Human Team B	0.182893	0.538934	0.273105
System C	0.271074	0.458341	0.340669
System D	0.294731	0.52713	0.378072

The d′ values for Teams A and B did not differ significantly ($t = 1.19$, $p = > 0.05$). The other assessments did differ significantly from one another (t: A vs. C: 25.14, A vs. D: 39.85, B vs. C: 27.44, B vs. D: 42.51, C vs. D: 135.94). By comparison, the adjudicated reviews combining Team A and Team B judgments with that of a senior attorney showed a d′ of 0.835.

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. F_1 is a summary measure combining precision and recall. It is calculated according to the formula used in TREC Legal 2008:

$$F_1 = \frac{2Pr \times R}{Pr + R}$$

where Pr = precision and R = recall.

These scores are comparable to those obtained in TREC Legal 2008. In that study, the median precision was 0.27 and median recall was 0.21.

Discussion

This study is an experimental investigation of how well computer-aided systems can do relative to traditional human review. It is an elaboration and extension of the kind of research done under the auspices of the TREC Legal Track. Both projects are concerned with identifying processes and methods that can help the legal community to meet its discovery obligations.

Although the volume of information that must be processed during litigation continues to grow, the legal profession's means for dealing with that information is on the verge of collapse. The same techniques that worked 20 years ago, when electronically stored information was relatively rare, do not continue to provide adequate or cost-effective results today, when electronic discovery matters can extend to many terabytes of data.

According to the Federal Rules of Civil Procedure (Rule 26(g)), each party must certify at the end of the discovery process that their production has been complete and accurate after a reasonable enquiry. There can be disagreement about what constitutes a reasonable enquiry, but it would seem that, all other things being equal, one that does as well as traditional practice would be likely to be considered reasonable.

Accuracy and Agreement

In the ideal case, we would like to know how accurate each classification is. Ultimately, measurement of accuracy implies that we have some reliable ground truth or gold standard against which to compare the classifier, but such a standard is generally lacking for measures of information retrieval in general and for legal discovery in particular. In place of a perfect standard, it is common to use an exhaustive set of judgments done by an expert set of reviewers as the standard (e.g., as is the practice in the TREC studies).

Under these circumstances, agreement with the standard is used as the best available measure of accuracy, but its acceptance should be tempered with the knowledge that this standard is not perfect.

Variability of Human Relevance Judgments

The level of agreement among human reviewers is not strikingly high. The two re-review teams agreed with the original review on about 76% and 72% of the documents. They agreed with one another on about 70% of the documents with corresponding kappa values in the low to fair range. Although low, these levels are realistic. They are comparable to those observed in the TREC studies and other studies of interrater agreement (e.g., Barnett et al., 2009, Borko, 1964; van Rijsbergen, 1979; Tonta, 1991; Voorhees, 1998).

There are two sources of this variability. Some variability is due to random factors, that is, factors that are unrelated to the material being judged or to any stable trait of the judges. For example, reviewers' attention may wander, they may be distracted, or fatigued. A document that they might have categorized as responsive when they were more attentive might then be categorized as nonresponsive or vice versa.

The second source of variability is systematic, which is due to the interaction between the content of the documents and stable properties of the reviewers, and to individual differences among reviewers.

Relevance judgments may be strategic. Reviewers may have different goals in mind when assessing documents and these goals may vary over time. Differences in strategic judgment may affect how likely two individuals are to call a certain document responsive. As noted by the TREC Legal 2008 Topic Authorities (http://trec-legal.umiacs.umd.edu/TAreflections2008.doc, retrieved May 7, 2009):

> While the ultimate determination of responsiveness (and whether or not to produce a given document) is a binary decision, the breadth or narrowness with which "responsiveness" is defined is often dependent on numerous subjective determinations involving, among other things, the nature of the risk posed by production, the party requesting the information, the willingness of the producing party to face a challenge for underproduction, and the level of knowledge that the producing party has about the matter at a particular point in time. Lawyers can and do draw these lines differently for different types of opponents, on different matters, and at different times on the same matter. This makes it exceedingly difficult to establish a "gold standard" against which to measure relevance/responsiveness and explains why document review cannot be completely automated.

Instead of "subjective," it may be more appropriate to say that discovery involves judgment about the situation as well as about the documents and their contents. Some judgments bias the reviewer to be more inclusive and some bias the reviewer to be less inclusive, but these judgments are not made willy-nilly. As opposed to pure errors, which are random, these judgment calls are based on a systematic interpretation of the evidence and the situation. To the extent that judgments are systematically related to the content of the documents, even if biased, they are capable of being mirrored by some automated system. The classifications made by an automated system can easily include the bias judgments of the attorneys managing a case, being either more or less inclusive as the situation warrants. Bias is not a barrier to automation, despite the implication drawn by the TREC Legal Topic Authorities.

Nevertheless, bias can change from case to case and individual to individual. It is not a stable property of the methods used to categorize the documents, so it is helpful to distinguish the power of the method from the bias to be more or less inclusive. Signal detection theory, by separating bias from discriminability, allows us to recognize the role of the information in the document contents and the sensitivity of the method. The d' values observed in the study showed that the human reviewers were no better at distinguishing responsive from nonresponsive documents than were the two automated systems.

Discovery cannot be wholly automated, not for the reason that it involves so-called subjective judgment, but because ultimately attorneys and parties in the case have to know what the data are about. They have to formulate and respond to arguments and develop a strategy for winning the case. They have to understand the evidence that they have available and be able to refute contrary evidence. All of this takes knowledge of the case, the law, and much more.

When judgments are made by review teams, they necessarily add to the variability of these judgments. Of the 225 attorneys conducting the review, few if any of them had much detailed knowledge of the business issues being considered, the case strategy, or the relative consequences of producing more or fewer documents before embarking on their review. There were certainly individual differences among them. Some of them were almost certainly better able to distinguish responsive from nonresponsive documents. And, moreover, the long arduous hours spent reviewing documents almost certainly resulted in fatigue and inattention. All of this variability does not lead to the creation of a very solid standard against which to compare other approaches to review. On the other side, the procedure of using many attorneys to conduct a review is current practice in large cases, so these results represent a realistic if not particularly reliable standard.

Anything that reduces this variability is likely to improve the level of agreement. One reason that recall rates are so low in the TREC Legal studies (and in the present study) is because of nonsystematic variability in the judgments that are being used as the ground truth. Reducing that variability, as the TREC Interactive Task did, improved recall by as much as 47% (Topic 103, H5). Similar factors are undoubtedly operating in this study. Adjudication, for example, improved the agreement between the combined judgments of Teams A and B with the original review. These differences again show the effect of bias. Teams A and B classified more documents as responsive than appeared in the adjudicated results. Using TREC methodology, this difference would show up as a decline in recall and an increase in precision with adjudication. Both the original review and the two human re-reviews reflected variable judgments.

Conversely, when we reduce the variability of one of the categorizers, in this case by using computer software to implement the judgments, then it may be possible to improve the measured level of agreement, even when compared to a variable standard. A given person may make different decisions about the same text at different times, while computer classifiers generally make consistent judgments. Comparing the decisions made by two variable processes is likely to lead to lower observed levels of agreement than would comparing a variable process to an invariant one. If the computer does not contribute its own variability to the agreement measure, then higher levels of agreement may be observed.

Effects of Base Rate

Because of the difference in base rates of responsive and nonresponsive documents, we used several measures to reduce the influence of simple chance on our measures. If high levels of agreement or accuracy were achieved simply because of base-rate differences, then separating the measures into positive and negative agreement would eliminate these differences. Even when eliminating differences in base rate by comparing within category, positive and negative agreement both show the same pattern of results.

As another approach to assessing agreement independent of base-rate differences, two chance-corrected measures, kappa and phi, were also used. Systems C and D showed at least as high a level of agreement on these measures as was found using Team A and Team B.

Blair and Maron Revisited

Blair and Maron (1985) found that their attorneys were able to find only about 20% of the responsive documents. They concluded that it was impossibly difficult to guess the right words to search for and instead advocated for using human indexers to develop a controlled vocabulary. Collections that seemed large to Blair and Maron, however, are dwarfed by the size of the present collection and many collections typical of modern electronic discovery. Employing human reviewers to manually categorize the documents can cost millions of dollars, an expense that litigants would prefer to reduce if possible.

Blair and Maron argued for using human readers to assign documents to specific categories because, they concluded, guessing the right terms to search for was too difficult to be practical. In contrast, with the size of modern collections, lawyers are finding that human categorization is too expensive to be practical.

The categorization systems used in the present study, and many others in current use, are more elaborate than the search system used by Blair and Maron. They employ more information about the documents and the collection as well as information from outside the collection (such as an ontology or the results of human classification). Many of these elaborations are designed to overcome the problem of guessing query terms.

Our best estimates from the present study suggest that both human review teams and computer systems identified a higher percentage of responsive documents than Blair and Maron's participants did. It is interesting to note that the human reviewers of Teams A and B were not more successful than the computer systems were at identifying responsive documents. One limitation may be the variability of the human judgments against which the computer systems are being compared.

Comparison With TREC Legal

The results of this study are generally congruent with those produced by TREC Legal. The methodology used in the present study has some advantages and some disadvantages relative to that used by TREC, but the differences typically are more indicative of the difficulty of doing this kind of research than of any flaw in design. They are predominantly responses to constraints, not errors.

By its charter, TREC is required to use publicly available datasets. Realistic litigation data, in contrast, are typically highly confidential and difficult to obtain for research purposes. For its first 3 years of investigations, TREC concentrated on a large collection of tobacco-related documents that were released as part of a legal settlement. These documents were mostly converted into electronic text using optical character recognition (OCR), which introduces errors. Because the documents in the collection were produced as part of a case, many of the irrelevant nonresponsive documents that are typical of actual electronic discovery collections were eliminated. Every document was deemed responsive to something. The TREC Legal designers have compensated for this by inventing issues/topics that might have been litigated. Their performance measures are based on sampling.

The present study, in contrast, used a real matter based on a Department of Justice request for information about a merger. Therefore, the responsiveness categorization is more naturalistic. It would be preferable, perhaps, if the matter were a litigation rather than a DOJ request, but these are the data that were made available. On the other hand, these data have not been made publicly available. Although some documents (600,000 out of 2.3 million) were scanned and OCRed, the majority were native electronic documents. Rather than sampling, the original collection was exhaustively reviewed at substantial expense in the context of a legal matter without any plans, at the time, for conducting a study. It would be very difficult to replicate this exhaustive review as part of a research project.

The reviews in the present study were performed by attorneys; in the TREC Legal studies the reviewers were predominantly law students. In the present study the reviewers spent hundreds of hours reviewing documents under some time pressure; in the TREC Legal study each reviewer spent about 21 hours reviewing documents at their own pace.

Another difference between the present study and the TREC Legal study is the use of documents that are more typical of modern electronic discovery situations than were many of the Tobacco documents. A majority of the documents in the present study (1.5 million) consisted of emails. The Tobacco collection contains a smaller proportion of emails, consisting rather of internal memos and other documents (Eichman & Chin, 2007).

In TREC Legal, many of the human assessor–assessor relations were computed on relatively small numbers of documents and typically involved equal numbers of responsive or nonresponsive documents. Decision bias is known to be affected by the proportion of positive events (e.g., Green & Swets, 1966). In contrast, the present study used naturalistic distributions of responsive and nonresponsive documents and larger sample sizes for the comparison of assessor–assessor relations. Still, both studies found similar levels of agreement.

Finally, the present study used two commercial electronic discovery service providers, whereas TREC is open to anyone who wants to contribute. These providers volunteered their processing time and effort to categorize the data. Although a few active service providers contributed to the TREC results, most of the contributors were academic institutions, so it is difficult to generalize from the overall performance of the TREC Legal participants to what one might expect in electronic discovery practice. Academic groups might be either more or less successful than commercial electronic discovery organizations.

The results from each service provider in the present study are displayed anonymously. These volunteers were intended to be representative of the many that are available. With the large number of documents involved, any slight difference between them is likely to be statistically significant, but small differences are not likely to be meaningful or replicable. The goal was to determine whether these tools could provide results comparable to those obtained through a complete manual review, and in that they have succeeded.

Conclusion

This study is an empirical assessment of two methods for identifying responsive documents. It set out to answer the question of whether there was a benefit to engaging a traditional human review or whether computer systems could be relied on to produce comparable results.

On every measure, the performance of the two computer systems was at least as accurate (measured against the original review) as that of a human re-review. Redoing the same review with more traditional methods as was done during the re-review had no discernible benefit.

There may be other factors at play in determining legal reasonableness, but all other things being equal, it would appear that employing a system like one of the two systems employed in this task will yield results that are comparable to the traditional practice in discovery and would therefore appear to be reasonable.

The use of the kind of processes employed by the two systems in the present study can help attorneys to meet the requirements of Rule 1 of the Federal Rules of Civil Procedure: "to secure the just, speedy, and inexpensive determination of every action and proceeding."

References

Bace, J. (2007). Cost of e-discovery threatens to skew justice system. Gartner Report G00148170. Retrieved May 6, 2009, from http://www.akershaw.com/Documents/cost_of_ediscovery_threatens_148170.pdf

Baron, J.R. (2008). Beyond keywords: Emerging best practices in the area of search and information retrieval. New Mexico Digital Preservation Conference, June 5, 2008. Retrieved July 29, 2009, from http://www.archives.gov/rocky-mountain/records-mgmt/conferences/digital-preservation/beyond-keywords.pdf

Baron, J.R., Lewis, D.D., & Oard, D.W. (2007). TREC-2006 Legal Track Overview. In Proceedings of the 15th Text REtrieval Conference (TREC 2006) (pp. 79–99). Gaithersburg, MD: NIST. Retrieved September 21, 2009, from http://trec.nist.gov/pubs/trec15/papers/LEGAL06.OVERVIEW.pdf

Barnett, T., Godjevac, S., Renders, J.-M., Privault, C., Schneider, J., & Wickstrom, R. (2009, June). Machine learning classification for document review. Paper presented at the ICAIL 2009 Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery. Retrieved July 24, 2009, from http://www.law.pitt.edu/DESI3_Workshop/Papers/DESI_III.Xerox_Barnett.Xerox.pdf

Blair, D.C., & Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM, 28, 289–299.

Borko, H. (1964) Measuring the reliability of subject classification by men and machines. American Documentation, 15(4), 268–273.

Eichman, D., & Chin, S.-C. (2007, June). Concepts, semantics and syntax in e-Discovery. Paper presented at the ICAIL 2007 Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery. Retrieved July 24, 2009, from http://www.umiacs.umd.edu/~oard/desi-ws/papers/eichmann.pdf

Green, D.M., & Swets, J.A. (1966). Signal detection theory and psychophysics. New York: John Wiley & Sons.

Gruner, R.H. (2008). Anatomy of a lawsuit. Retrieved July 24, 2009, from http://www.vallexfund.com/download/AnatomyLawsuit.pdf

Oard, D.W., Hedin, B., Tomlinson, S., & Baron, J.R. (2009). Overview of the TREC 2008 Legal Track. In Proceedings of the 17th Text Retrieval Conference (TREC 2008). Retrieved September 21, 2009, from http://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf

Paul. G.L., & Baron, J.R. (2007). Information inflation: Can the legal system adapt? Richmond Journal of Law and Technology, 13, Article 10, 1–41. Retrieved July 28, 2009, from http://law.richmond.edu/jolt/v13i3/article10.pdf

Rijsbergen, C.J. van (1979). Information retrieval, 2nd ed. London: Butterworths.

Swets, J.A. (1969). Effectiveness of information retrieval methods. American Documentation, 20(1), 72–89.

Tomlinson, S., Oard, D.W., Baron, J.R., & Thompson P. (2008). Overview of the TREC 2007 Legal Track. In Proceedings of the 16th Text REtrieval Conference (TREC 2007). Retrieved September 21, 2009, from http://trec.nist.gov/pubs/trec16/papers/LEGAL.OVERVIEW16.pdf

Tonta, Y. (1991). A study of indexing consistency between Library of Congress and British Library catalogers, Library Resources & Technical Services, 35(2), 177–185.

Voorhees, E.M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 315–323). New York: ACM Press.