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,303 South Broadway, Suite 430,Tarrytown, NY 10591. E-mail: patrick@ediscoveryinstitute.org

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

 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

DOI: 10.1002/asi

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; Table 2, 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

DOI: 10.1002/asi

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 $\sim$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

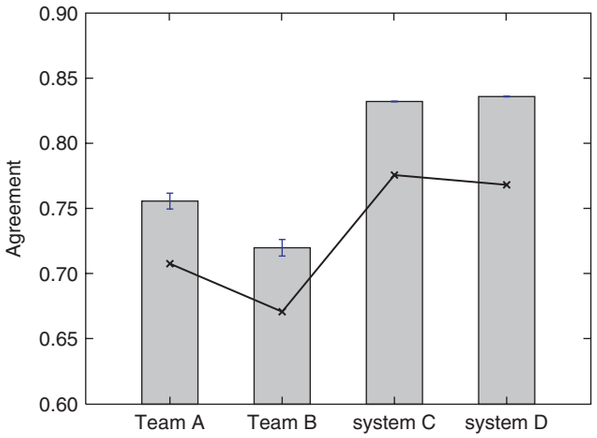

FIG. 1. The level of agreement with the original review and chance levels to be expected from the marginals for the two human teams and the two computer systems (the four reassessments). Error bars show standard error.

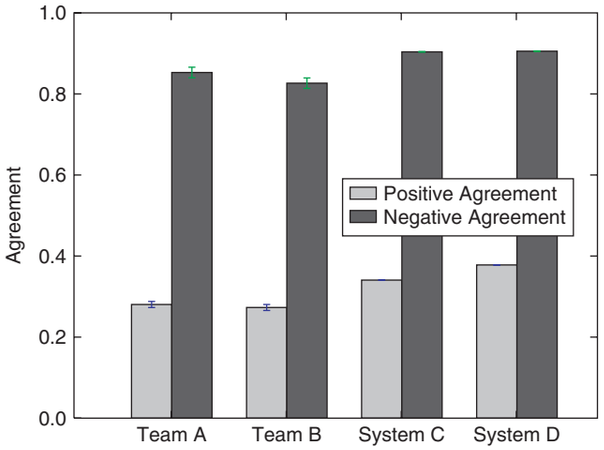

FIG. 2. Positive agreement (2*RR/(2*RR + RN + NR)) and negative agreement (2*NN/(2*NN + NR + RN)) for agreement between the original review and the four reassessments. NN, NR, etc. refer to the columns of Table 1. Error bars are standard error.

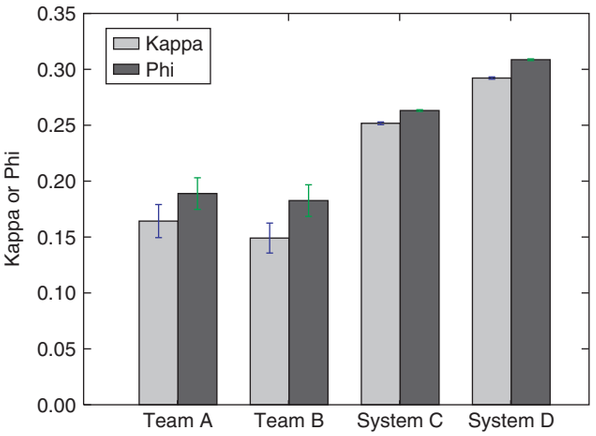

FIG. 3. Kappa and Phi for the agreement between the original review and the four reassessments. Kappa and Phi are "chance adjusted" measures of association or agreement. Error bars are standard error.

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 = .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 = .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 $p < .01$).

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).

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

DOI: 10.1002/asi

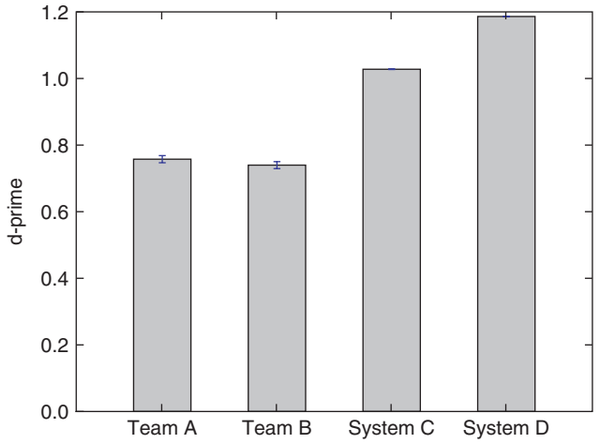

FIG. 4. The signal detection measure d′ comparing each of the re-reviews against the original review.

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

DOI: 10.1002/asi

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

DOI: 10.1002/asi

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

DOI: 10.1002/asi

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.

DOI: 10.1002/asi