Re-examining the Effectiveness of Manual Review

William Webber

Computer Science and Software Engineering
The University of Melbourne
Victoria 3010, Australia
wew@csse.unimelb.edu.au

ABSTRACT

Recent studies have found that automated retrieval methods in e-discovery are not only cheaper than manual review, but are also as or more reliable. We argue that these studies, while suggestive, are not conclusive. There is a high variability in the quality of unsupervised manual reviewers, as data from the TREC Legal Track shows. The best manual reviewers are as reliable as automated methods, and a properly supervised manual review may be more reliable than automation. We show the effectiveness of the simple review management approach of observing the proportions found relevant between reviewers. Finally, we describe the experimental protocol necessary for a more conclusive comparison of manual and automated review.

1. INTRODUCTION

The volume of electronically-stored information (ESI) held by modern corporations is driving discovery to use various forms of technology-assisted or automated review. An important question is whether automated methods are merely a cheaper but lower-quality alternative to full manual review, or whether automation leads to document productions of equal or even higher quality. The former alternative means automation is a compromise; the latter would make full manual review obsolete.

Two recent studies have compared the quality of automated retrieval and manual review, one by a re-review of an earlier manual production [Roitblat et al., 2010], the other through an analysis of data from the TREC 2009 Legal Track [Grossman and Cormack, 2011]. The former study finds automated retrieval to be at least as consistent as manual review, while the latter concludes that automation gives superior reliability.

We revisit the comparison of automated and manual review methods, and argue that the previous studies, though suggestive, are not conclusive. In particular, we re-examine the TREC Legal Track data, observing that the reviewers used are of highly variable reliability. The best reviewers are of comparable or better quality than the best automated systems, even under the asymmetric experimental conditions of the track. It is still open to question, therefore, whether an automated system can surpass or even achieve the reliability of a properly managed manual review team.

Whether automated tools have surpassed manual review in quality is a question too important to leave without a firm answer. We therefore conclude our paper with what is required for an experimental program to answer this question more conclusively.

SIRE'11, July 28, 2011, Beijing.
ACM SIGIR.

2. BACKGROUND

It is well known that human assessors frequently disagree on the relevance of a document to a topic. Voorhees [2000] found that experienced TREC assessors, albeit working from only sentence-length topic descriptions, had an average overlap (size of intersection divided by size of union) of between 40% and 50% on the documents they judged to be relevant. Voorhees concludes that 65% recall at 65% precision is the best retrieval effectiveness achievable, given the inherent uncertainty in human judgments of relevance. Bailey et al. [2008] survey other studies giving similar levels of inter-assessor agreement.

When one conception of relevance is authoritative, assessors do not merely disagree; they make errors. In legal discovery, the authoritative conception of relevance is that of the attorney overseeing the retrieval. The Interactive Task of the Legal Track of TREC includes such a *topic authority*, and provides a process of appeal to this authority for uncovering assessor errors (Section 2.3). The appeal results for TREC 2009 found that, on an assessment set in which 90% of documents were actually irrelevant, 33% of relevant assessments were in error, as were 3% of irrelevant assessments [Hedin et al., 2009]. This is likely a lower bound to the error rate, since some errors may not have been appealed (although conversely some appeals may have been erroneously upheld).

Since assessors disagree, and reviewers make mistakes, the production of a manual review process is not an inerrant gold standard, which an automated process might approach but cannot surpass. The question, rather, is whether a manual or an automated review process gives more reliable results. That is the topic addressed by the studies described below.

2.1 Terminology and measures

Manual review denotes a process in which every candidate document for production is reviewed for relevance by at least one human reviewer. Candidate documents might be every document in a corporation's possession, but generally some prior filtering has been performed, by custodian for instance, or by keyword queries, though the latter blurs the line between manual and automated review. Automated review denotes a situation in which the decision to produce or not produce some proportion of the candidate documents is made algorithmically, without complete human review. The term "technology-assisted review" is often used instead, but while this may be softer to a judge's ears, it seems to us inexact and unhelpful; surely all review of ESI requires the assistance of at least some degree of technology.

We use three measures of a retrieval's effectiveness: precision, recall, and the F1 score. Precision is the proportion of retrieved documents that are relevant; recall is the proportion of relevant documents that retrieved. There is a natural tension between the two

measures: shrinking the retrieved set generally helps precision, but can only decrease recall; expanding the retrieval can only help recall, but generally hurts precision. This tension is captured in the F1 measure, which is the harmonic mean of precision and recall:

$$F_1 = \frac{2 \cdot R \cdot P}{R + P} \ . \quad (1)$$

2.2 Roitblat, Kershaw, and Oot [2010]

The first study comparing manual and automated review that we consider is that of Roitblat et al. [2010]. For their study, the authors revisit the outcome of an earlier, in-house manual review. The original review surveyed a corpus of 2.3 million documents in response to a regulatory request, and produced 176,440 as responsive to the request; the process took four months and cost almost $14 million. Roitblat et al. had two automated systems and two manual review teams review the documents again for relevance to the original request. The automated systems worked on the entire corpus; the manual review teams looked at a sample of 5,000 documents.

Roitblat et al. (Table 1) found that the overlap between the relevance sets of the two manual teams was only 28%, even lower than the 40% to 50% observed in Voorhees [2000] for TREC AdHoc assessors. The overlap between the new and the original productions was also low, 16% for each of the manual teams, and 21% and 23% for the automatic systems. When compared against the original production, the human review teams achieved F1 scores of 0.27 and 0.28, while the automated systems achieved 0.34 and 0.38.

The effectiveness scores calculated on the original production seemingly show that the automated systems are as reliable as the manual reviewers. However, as Roitblat et al. note, the original production is a questionable gold standard, since it likely is subject to the same variability in human assessment that the study itself demonstrates. Instead, the claim Roitblat et al. make for automated review is a more cautious one; namely, that two manual reviews are no more likely to produce results consistent with each other than an automated review is with either of them.

Given the remarkably low level of agreement observed by Roitblat et al., their conclusion might seem a less than reassuring one; an attorney might ask not, which of these methods is superior, but, is either of these methods acceptable? More importantly, the study does not address the attorney's fundamental question: does automated or does manual review result in a production that more reliably meets the overseeing attorney's conception of relevance?

2.3 The TREC legal track

The Legal Track of TREC provides an objective environment in which to validate and compare different retrieval methods for e-discovery [Baron et al., 2006]. Since to date no participant has performed a fully manual review, there has not been a direct comparison of automated and manual review methods, though (as will be seen shortly) Grossman and Cormack [2011] present a method for extracting such a comparison from the TREC data.

Of particular interest for comparing manual and automated review is the track's Interactive Task. The task seeks (within experimental limits) to replicate the conditions of a real-world retrieval. In particular, there is a topic authority (TA), who plays the role of the attorney overseeing the production, and whose conception of relevance is authoritative. Teams may consult with the TA while producing their runs, and the TA instructs (though does not directly supervise) the track's relevance assessors. Teams may also appeal initial assessments to the TA for adjudication, with the adjudicated assessments forming the official assessment set for the task.

The dataset used by Grossman and Cormack [2011], and by the current paper, comes from the TREC 2009 Interactive Task. Seven topics were run that year; their assessment outcomes are summarized in Table 1. In the task's assessment scheme, messages are sampled from strata defined by participating team's intersecting productions, and also from the *bottom stratum* of messages returned by no system; the latter stratum is sampled sparsely, giving each sampled message a significant weight in effectiveness estimates [Hedin et al., 2009]. Documents (email bodies and attachments) in the messages sampled for assessment are assigned to sets called *bins* (column 2 of Table 1). Each bin is assessed by a single assessor; an assessor may (rarely) assess more than one bin. Most bins are core bins, to which messages are randomly assigned. A small number of supplementary bins, with differing assignment methods, are used to achieve special assessment goals.

Topic	Bins	Type	Ass'd	Ass Rel	Appl'd	Adj Rel
t201	13	Student	2729	328	305	195
t202	13	Student	3201	549	365	661
t203	12	Prof'nl	3320	113	254	225
t204	12	Prof'nl	3101	80	191	166
t205	12	Student	3002	1018	642	568
t206	12	Student	2770	130	34	99
t207	13	Prof'nl	2505	215	106	262

Table 1: Summary of assessment for the interactive task topics of the TREC 2009 Legal Track. Reported are number of core bins; assessor type (law student or professional reviewer); number of messages sampled and assessed; number of messages initially assessed relevant; number of assessments appealed; number of messages assessed relevant after adjudication.

Assessors were of two types in 2009 (column 3 of Table 1): first, volunteer law students; or second, professional manual reviewers. Each bin was assigned enough messages (summed in column 4) to make up 500 documents. The number of messages initially assessed relevant varies widely between topics (column 5), as does the number of appeals (column 6). Since appealing was at the discretion of the participating teams, the latter variety could be due either to the errors of the assessors, or to the thoroughness of the teams. How complete the appeals were in detecting errors in the initial assessments is considered in Section 3.1.

2.4 Grossman and Cormack [2011]

Grossman and Cormack [2011] re-analyze the interactive task as a comparative evaluation of manual and automated review, by treating the assessors as a manual review team, and evaluating their retrieval, alongside that of the automated systems, against the adjudicated assessments. They select for this comparison two top-performing automated systems: an industry system which we will name System I, and an academic one, System A. The five topics in which these systems participated were heavily appealed, in particular by these teams themselves, leading to good coverage of assessor errors—or, perhaps, a re-alignment of the TA's conception of relevance with the appealing team's.

The outcome of the evaluation performed by Grossman and Cormack is shown in Table 2. The automated systems beat the manual review teams quite handsomely for four of the five topics, and come close for the fifth. On this showing, automated retrieval appears not merely an adequate, but a superior, alternative to manual review.

The analysis of Grossman and Cormack assumes that the adjudicated assessments are a "reasonably accurate gold standard", in the authors' words. This in turn requires that the appeal process is

Topic	Team	Rec	Prec	F1
t201	System A	0.78	0.91	0.84
	TREC (Law Students)	0.76	0.05	0.09
t202	System A	0.67	0.88	0.76
	TREC (Law Students)	0.80	0.27	0.40
t203	System A	0.86	0.69	0.77
	TREC (Professionals)	0.25	0.12	0.17
t204	System I	0.76	0.84	0.80
	TREC (Professionals)	0.37	0.26	0.30
t207	System A	0.76	0.91	0.83
	TREC (Professionals)	0.79	0.89	0.84

Table 2: Automated and manual reviewer effectiveness. Evaluation is against the adjudicated assessments, extrapolated to the full corpus of messages. The best automated team for the selected topics is compared to the manual review team constructed from the initial assessments of the track assessors. (Based upon Table 7 of Grossman and Cormack [2011]; values are recalculated.)

both reasonably complete and unbiased. Incomplete appeals would leave assessment errors unfound, inflating the effectiveness of manual review. On the other hand, appeals could shift the topic authority's conception of relevance towards a team's run, especially since (unlike the original assessments) they are accompanied by written justifications. Which of these two effects is stronger is unclear. More importantly, the re-purposed assessments are not true manual review efforts. How representative they are of a properly supervised manual review is the topic of Section 3.

3. RECONSIDERING MANUAL REVIEW

The previous section surveyed two recent studies comparing the reliability of manual and automated review. Next, we re-examine the measurement of manual review effectiveness, looking in particular at the evidence provided by the TREC 2009 Legal Interactive task.

3.1 Completeness of appeal process

First, what evidence do we have for the completeness of the appeals process, assumed by [Grossman and Cormack, 2011]? Since messages sampled for assessment are randomly assigned to core assessment bins, we should expect each bin to have the same proportion of relevant messages, subject to random variation.[1] Unevenness in proportions initially assessed relevant is evidence of assessor errors, and continued unevenness after adjudication is evidence that the appeals process has failed to uncover all such errors. The converse is not necessarily true: proportions could be balanced same even if many assessor errors exist, though this would be likely in practice only if the assessors as a group had a consistent, though incorrect, conception of relevance.

We illustrate the analysis of proportions assessed relevant, taking Topic 201 as an example. Figure 1 shows the proportion of messages in each core bin for this topic that were assessed relevant,

[1]The cohesion would be even stronger if the assignment were performed so that each bin received the same proportion of documents from each stratum, but this latter step was not in fact enforced. Note that we rely on the simple random sampling of messages in our analysis, not of documents; the latter are not simple-randomly sampled, but are clustered by messages.

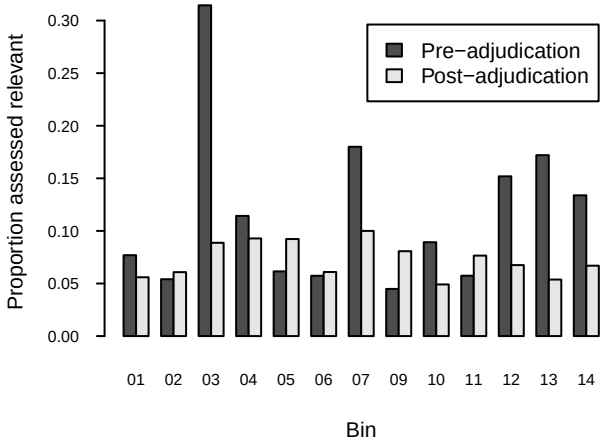


Figure 1: Proportion of messages assessed relevant in each core bin, prior to and after the appeal and adjudication process, for Topic 201.

Topic	Bins	Proportion χ^2	
		Assessed	Adjudicated
t201	13	150.7 **	9.9
t202	13	48.9 **	14.7
t203	12	68.3 **	11.1
t204	12	39.7 **	9.9
t205	12	367.0 **	45.5 **
t206	12	335.7 **	235.5 **
t207	13	10.1	8.1

Table 3: Chi-squared statistic for proportions relevant across core bins before and after adjudication. Proportions significantly uneven at $p < 0.001$ are marked with "".**

before and after adjudication. The proportions relevant in the initial assessment vary enormously; the mean proportion is 12%, but one assessor finds 31% of messages relevant, another just 4%. This provides clear evidence of many assessor errors. On the other hand, after appeal and adjudication, the mean proportion falls to 7%, and the range shrinks to between 5% and 10%.

We can test whether the relevance proportion between bins is uneven to a statistically significant degree using a χ^2 test of proportions. The null hypothesis is that reviewers are applying exactly the same conception of relevance, and that variability in proportions assessed relevant is due to sampling error alone. The χ^2 statistic measures the ratio between the observed and the expected variability between proportions (subject to the number of observations). The two-tailed expected 95% range of χ^2 for 12 bins is 3.8 to 21.9, for 13 bins 4.4 to 23.4. Values above that range indicate significant unevenness; values below would indicate suspicious evenness (suggesting, for instance, that teams set out to produce the same proportion relevant per bin, regardless of actual relevance).

The observed χ^2 statistics for the TREC 2009 topics, before and after adjudication, are given in Table 3. Prior to adjudication, the assessments for all topics other than Topic 207 show highly significant degrees of unevenness in proportions relevant between bins. After adjudication, five topics are not significantly uneven, being the five topics examined by Grossman and Cormack [2011]. The

appeal process appears to have been reasonably complete for these topics. Topic 206 was only lightly appealed, as Table 1 indicates, and highly significant unevenness remains; we can therefore regard that topic's adjudicated assessments as a poor gold standard, and exclude the topic from further analysis. Topic 205, in contrast, was the most heavily appealed topic, and yet significant unevenness remains; either there were an extraordinary number of assessor errors, or something untoward has occurred with the assessment process. Still, the degree of unevenness is greatly reduced through adjudication; we retain this topic in our subsequent analyses.

The proportions relevant of Topic 207's professional review team show the expected degree of evenness even before adjudication. That this evenness is evidence of a good review process is shown by the high reliability the team achieves in the analysis of Grossman and Cormack [2011] (Table 2), and is further confirmed by the examination of the reliability of individual assessors, below. The potential of the simple statistical analysis of evenness between proportions as a tool for review process control is examined later.

3.2 Sample and population accuracy

Some of the manual reviewer reliability figures given in Table 2 are rather alarming; for instance, that the review team for Topic 201 achieved a precision of only 0.05, returning only one actually relevant message in every twenty they judged relevant. This is not the reliability observed on the messages actually sampled, though; rather, it is the reliability extrapolated to the full population. Unequal sampling emphasises bottom stratum assessments overturned on appeal. For instance, for Topic 201, from one in two to one in eight messages were sampled from upper strata, but only one in three-hundred from the bottom stratum. Each successful appeal carries up to 150 times the weight on the bottom stratum that it does on the upper ones. Of the 1,927 messages sampled from the bottom stratum for this topic, 72 were found relevant by the assessors, but 71 of these assessments were appealed, and all 71 were overturned on appeal; this is why such low precision is reported for the reviewers in Table 2.

The strong weight on these bottom-stratum appeals means that even a slight appeal-induced bias would greatly harm the apparent precision of the reviewers, and boost the recall of the teams. Moreover, even if the figures are taken at face value, what is being simulated here is essentially an unsorted linear review of the full corpus, and the errors of (presumably) inattention that such a review would turn up. Such an exhaustive linear review might be prevented in practice by a pre-filtering by custodian or keyword; and errors of inattention would be readily picked up by dual-assessment, particularly of assessed-relevant messages.

For comparison with the extrapolated reliability figures in Table 2, we recalculate in Table 4 both team and reviewer accuracy on the post-adjudication sample of messages alone, without extrapolating to the full population. The relative ordering of team and reviewer is the same as on the population (Table 2), with the best team better than the composite of reviewers for every topic except Topic 207. The performance of the weaker review teams, however, is less extreme than under extrapolation. For instance, the team of student reviewers for Topic 201 scored a precision of 0.05 and an F1 score of 0.09 on the population, due to 71 of their 72 relevance assessments on the sparsely-sampled bottom stratum being overturned on appeal; judged on the sample only, however, their precision improves to 0.41, and their F1 score to 0.52.

The extrapolated reliability figures in Table 2 are not simply wrong, nor are the sample figures in Table 4 simply correct. The raw reliability figures given in the former case, however, need to be treated with some caution, due to the magnifying effect on errors of sampling, the potential for appeal-induced adjudication bias, and the lack of simple quality-control mechanisms. The setup of the assessment may not be a fair representation of an actual manual review. Nevertheless, for the following analysis, we will use the reviewer reliability figures as extrapolated to the population.

Topic	Team	Rec	Prec	F1
t201	System A	0.96	0.91	0.94
	TREC (Law Students)	0.70	0.41	0.52
t202	System A	0.81	0.88	0.84
	TREC (Law Students)	0.76	0.91	0.82
t203	System A	0.81	0.71	0.76
	TREC (Professionals)	0.25	0.50	0.34
t204	System I	0.94	0.84	0.89
	TREC (Professionals)	0.24	0.55	0.33
t207	System A	0.78	0.90	0.84
	TREC (Professionals)	0.78	0.93	0.85

Table 4: Automated and manual reviewer effectiveness, evaluated on the sampled assessments directly, without extrapolation to the full corpus. Other details are as for Table 2.

3.3 Variability in reviewer reliability

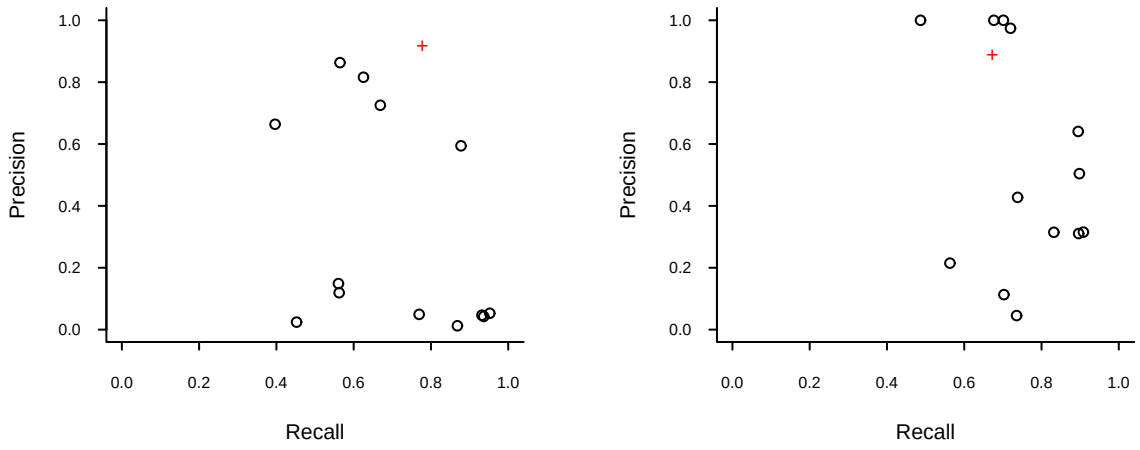
The reviewer reliability scores in Table 2 are averages across each team of assessors. Figure 1 and Table 3 indicate that for most assessment teams, there is great variability in the proportion of messages that each assessor finds relevant, which suggests that there may be similar variability in error rates. In this section, we directly investigate variability in assessor reliability.

Figure 2 shows the reliability of the review performed in each bin, evaluated against the adjudicated assessments, and compares it to the performance of the automated systems identified by Grossman and Cormack [2011]. For every topic but one (Topic 207), there is a great diversity between the reliability of different reviewers. Per-bin precision ranges from almost 0.0 to approaching 1.0, and the range of recall values is often 0.6 wide. Only for Topic 203 does the best automated system clearly outperform the best manual reviewer. As before, the professional manual review team for Topic 207 stands out. Several reviewers outperform the best automated system, and even the weaker individual reviewers have both precision and recall above 0.5.

The variability in reviewer reliability seen in Figure 2 suggests the importance of a proper review management process. The best reviewers generally match the best automated systems, even amongst student reviewers. A process that brought all reviewers up to the standards of the best performers, such perhaps as the process employed by the group in Topic 207, would seem to have the potential to offer equal or superior reliability to the best automated methods. Just excluding the weaker reviewers would by itself significantly improve review team reliability. The next section explores a simple mechanism for achieving this.

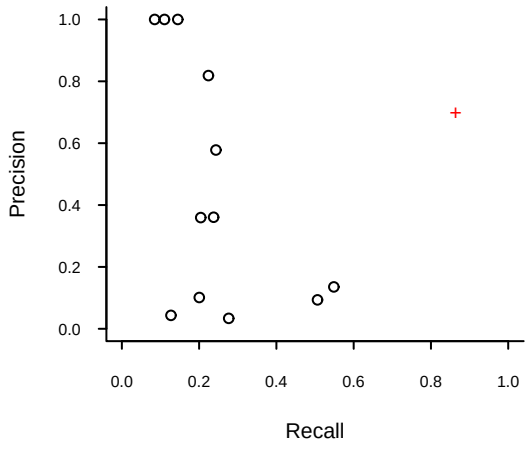
3.4 Improving review team quality

There are many tools that can be employed to improve the quality of a review process, some to do with human factors, others involving statistics. Dual assessment, for instance, can help catch random errors of inattention, while second review by an authoritative reviewer such as the supervising attorney can correct misconceptions of relevance during the review process, and adjust for

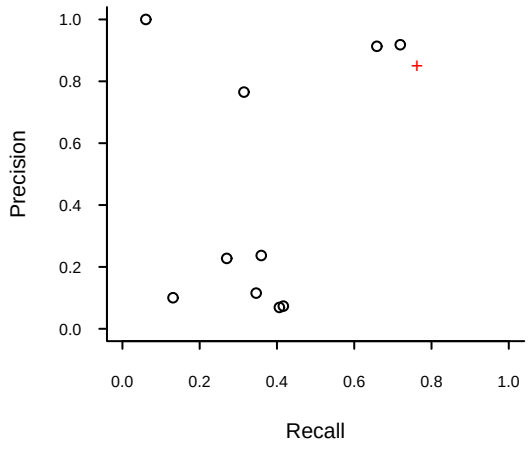


(a) Topic 201

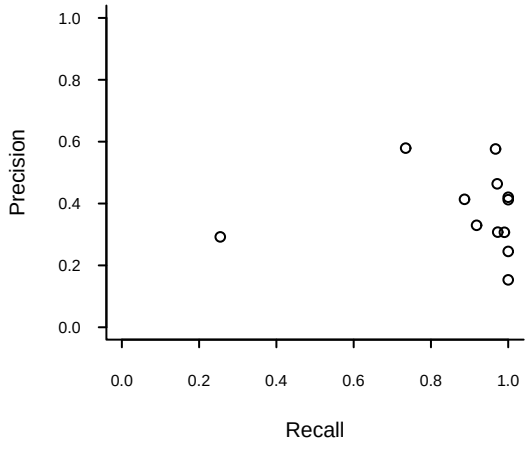
(b) Topic 202



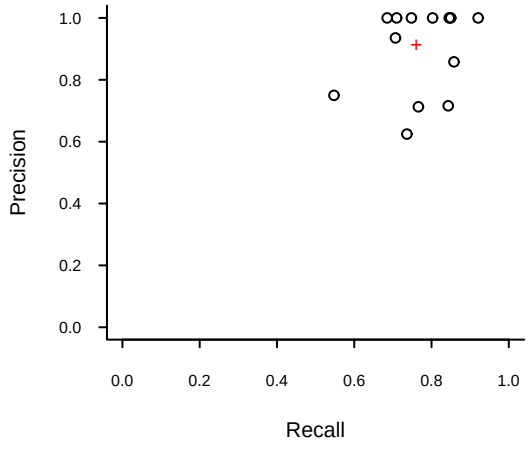
(c) Topic 203



(d) Topic 204



(e) Topic 205



(f) Topic 207

Figure 2: Assessor precision and recall, extrapolated to population, for Topics 201–205 and Topic 207. Each circle represents the reliability of a core bin. The red cross in each figure except that for Topic 205 gives the performance of the best automated retrieval effort, as listed in Table 2.

Topic	Reviewers	Rec	Prec	F1
t201	All	0.76	0.05	0.09
	Excl	0.59	0.12	0.19
t202	All	0.79	0.25	0.38
	Excl	0.81	0.38	0.52
t203	All	0.25	0.12	0.17
	Excl	0.18	0.25	0.21
t204	All	0.38	0.28	0.32
	Excl	0.47	0.28	0.35
t205	All	0.94	0.33	0.49
	Excl	0.97	0.43	0.60
t207	All	0.77	0.87	0.82
	Excl	0.77	0.87	0.82

Table 5: Review team effectiveness, including and excluding reviewers with a disproportionate number of relevant documents

assessor errors once it is complete [Webber et al., 2010]. Exploring the full range of process quality management techniques is beyond our current scope. It has already been observed, though, that the proportion assessed relevant is a simple indicator of overall review consistency and quality. Does it also indicate individual reliability, pointing out unreliable assessors for retraining or exclusion?

We begin by relating the proportion found relevant in a bin with the reliability of that bin, as measured by F1 score. The goal is to identify bins that are outliers in the proportion of messages they find relevant. To do this, we take the median proportion relevant across all bins (since the median is more robust to outliers than the mean), and determine which bins produce relevance proportions that are significantly different from the median, at level $p < 0.01$ in a two-tailed exact binomial test.

Figure 3 compares the pre-adjudication bin proportions relevant with F1 scores across the different TREC topics, indicating which bins are significantly different from the median proportion relevant. Note, first, the spread in proportions relevant, particularly the remarkable dispersion for Topic 205, revealing a review process that was clearly not in control. The relationship is not unanimous, but the more reliable bins tend to be those closer to the median proportion of messages relevant. In particular, significant divergence from the median appears to be a partial, though not infallible, indicator of reviewer unreliability.

A simple approach to improving review team quality is to exclude those reviewers whose proportion relevant are significantly different from the median, and re-apportion their work to the more reliable reviewers. Table 5 reports the change in review team reliability if this step is taken, considering only the documents falling into the non-excluded bins (or, equivalently, assuming the work from the excluded reviewers is re-apportioned evenly and performed to the same standard as the rest of each reviewer's bin). In accordance with our previous observations, there is a general improvement in reliability, though not always a great one. For every topic in which a bin is excluded (every topic, that is, except for the consistently-reliable Topic 207), the F1 score of the post-exclusion review team is higher than that of the original, sometimes by an appreciable margin. Precision also generally rises, but in a couple of cases recall falls, reflecting the fact that being overly generous in one's assessments can help draw in relevant documents one might otherwise have missed.

Fully excluding reviewers based solely on the proportion of documents they find relevant is a crude technique. Nevertheless, the results of this section suggest that this proportion is a useful, if only partial, indicator of reliability, one which could be combined with additional evidence to alert review managers when their review process is diverging from a controlled state. It may be that review teams with better processes, such at the team from Topic 207, already use such techniques. Therefore, they need to be considered when a benchmark for manual review quality is being established, against which automatic techniques can be compared.

4. ASSESSING REVIEW METHODS

Roitblat et al. and Grossman and Cormack have presented evidence for the equal or greater reliability of automated compared to full manual review. The former study, though, takes a manual review itself as the gold standard. The TREC experiments reanalyzed by Grossman and Cormack do use a human topic authority to measure production quality, something which is more representative of professional practice. We have observed in the Section 3, however, that the manual review pseudo-teams formed by re-purposing the track assessors are highly variable in quality, suggesting a lack of the quality control and direction that might be expected in a true, professional review effort.

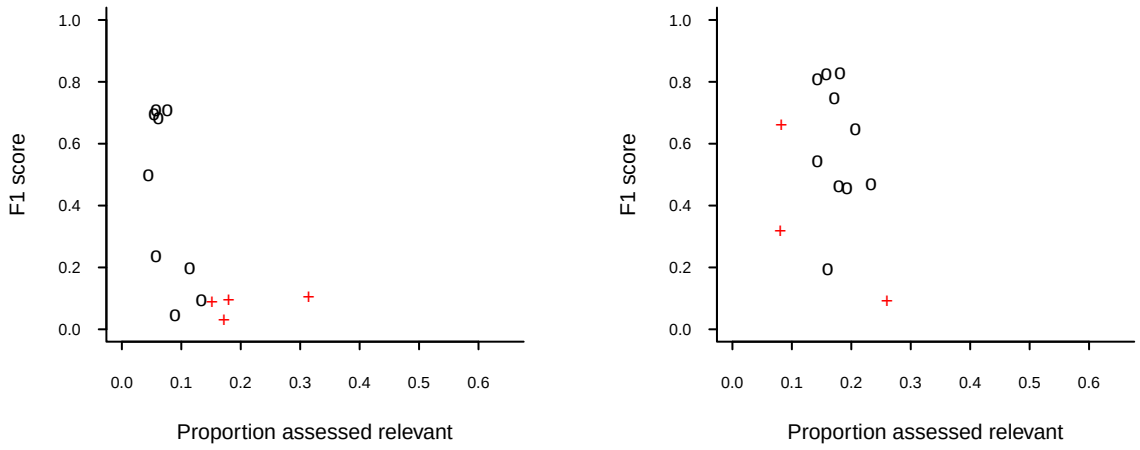
What is needed are experiments comparing automated and manual approaches on an even footing (as in Roitblat et al.), evaluated against the objective standard of a supervising topic authority (as in Grossman and Cormack). The authority should drive both productions, on the one topic: providing coding standards and supervision to the manual team, and seed queries and relevance assessments to the automated one. Both processes, particularly the manual review, should be conducted according to industry standards. The same topic authority should then assess the quality of each production, both for conformity to their own conception of relevance, and for the amount of effort involved in the production.

No single experiment of this sort can be comprehensive: there are a variety not only of automated review methods, but also of manual process strategies; and, of course, there are a multitude of potential corpora and production requests. And even such a setup as this involves a degree of unrealism and artificiality, since actual productions are made in several, possibly iterated, stages (extracting, culling, reviewing, redacting, collating), and inevitably with a complex mix of manual and automated processes. Nevertheless, such experiments, by directly comparing the two approaches on an equal footing, in a more realistic environment, and against a representative objective standard, will allow us to draw firmer conclusions on the relative merits of the manual and automated review.

5. CONCLUSIONS

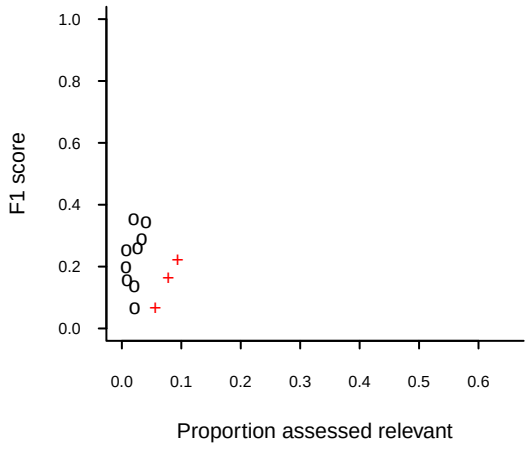
The original review from which Roitblat et al. draw their data cost $14 million, and took four months of 100-hour weeks to complete. The cost, effort, and delay underline the need for automated review techniques, provided they can be shown to be reliable. Given the strong disagreement between manual reviews, even some loss in review accuracy might be acceptable for the efficiency gained. If, though, automated methods can conclusively be demonstrated to be not just cheaper, but more reliable, than manual review, then the choice requires no hesitation. Moreover, such an achievement for automated text-processing technology would mark an epoch not just in the legal domain, but in the wider world.

Two recent studies have examined this question, and advanced evidence that automated retrieval is at least as consistent as manual review [Roitblat et al., 2010], and in fact seems to be more reli-

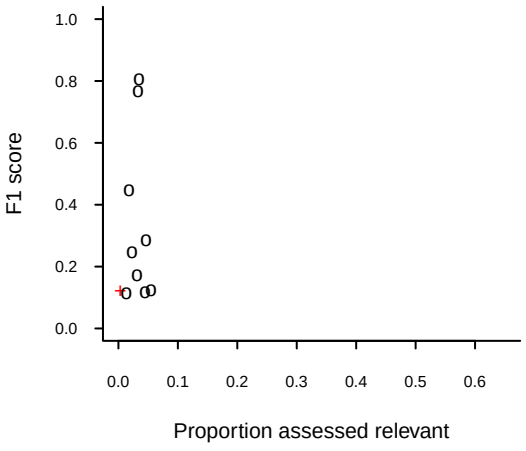


(a) Topic 201

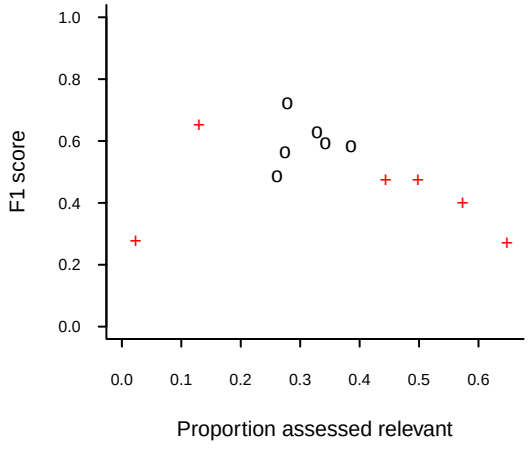
(b) Topic 202



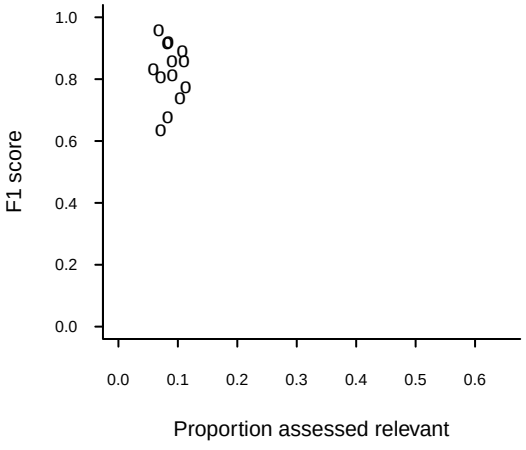
(c) Topic 203



(d) Topic 204



(e) Topic 205



(f) Topic 207

Figure 3: Assessor F1 score and proportion assessed relevant by bin for Topics 201–205 and Topic 207. Scores are extrapolated to the population; proportions assessed relevant are taken from the sample. Bins with a proportion relevant significantly different from the median ($p < 0.01$) are shown as red crosses; non-significant bins are black circles.

able [Grossman and Cormack, 2011]. These results are suggestive, but (we argue) not conclusive as they stand. For the latter study in particular (leaving questions of potential bias in the appeals process aside), it is questionable whether the assessment processes employed in the track truly are representative of a good quality manual review process.

We have provided evidence of the greatly varying quality of reviewers within each review team, indicating a lack of process control (unsurprising since for four of the seven topics the reviewers were not a genuine team). The best manual reviewers were found to be as good as the best automated systems, even with the asymmetry in the evaluation setup. The one, professional team that does manage greater internal consistency in their assessors is also the one team that, as group, outperforms the best automated method. We have also pointed out a simple, statistically based method for improving process control, by observing the proportion of documents found relevant by each assessor, and counselling or excluding those who appear to be outliers.

Above all, it seems that previous studies (and this one, too) have not directly addressed the crucial question, which is not how much different review methods agreed or disagree with each other (as in the study by Roitblat et al. [2010]), nor even how close automated or manual review methods turn out to have come to the topic authority's gold standard (as in the study by Grossman and Cormack [2011]). Rather, it is this: which method can a supervising attorney, actively involved in the process of production, most reliably employ to achieve their overriding goal, to create a production consistent with their conception of relevance. There is good, though (we argue) so far inconclusive, evidence that an automated method of production can be as reliable a means to this end as a (much more expensive) full manual review. Quantifying the tradeoff between manual effort and automation, and validating protocols for verifying the correctness of either approach in practice, are particularly relevant in the multi-stage, hybrid work-flows of contemporary legal review and production. Given the importance of the question, we believe that it merits the effort of a more conclusive empirical answer.

References

Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, A. de Vries, and Emine Yilmaz. Relevance assessment: are judges exchangeable and does it matter? In Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong, editors, *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 667–674, Singapore, Singapore, July 2008.

Jason R. Baron, David D. Lewis, and Douglas W. Oard. TREC-2006 legal track overview. In Ellen Voorhees and Lori P. Buckland, editors, *Proc. 15th Text REtrieval Conference*, pages 79–98, Gaithersburg, Maryland, USA, November 2006. NIST Special Publication 500-272.

Maura R. Grossman and Gordon V. Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, 17(3):11:1–48, 2011.

Bruce Hedin, Stephen Tomlinson, Jason R. Baron, and Douglas W. Oard. Overview of the TREC 2009 legal track. In Ellen Voorhees and Lori P. Buckland, editors, *Proc. 18th Text REtrieval Conference*, pages 1:4:1–40, Gaithersburg, Maryland, USA, November 2009. NIST Special Publication 500-278.

Herbert L. Roitblat, Anne Kershaw, and Patrick Oot. Document categorization in legal electronic discovery: computer classification vs. manual review. *Journal of the American Society for Information Science and Technology*, 61(1):70–80, 2010.

Ellen Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5): 697–716, September 2000.

William Webber, Douglas W. Oard, Falk Scholer, and Bruce Hedin. Assessor error in stratified evaluation. In *Proc. 19th ACM International Conference on Information and Knowledge Management*, pages 539–548, Toronto, Canada, October 2010.