UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

MONIQUE DA SILVA MOORE, MARYELLEN O'DONOHUE, LAURIE MAYERS, HEATHER PIERCE, and KATHERINE WILKINSON on behalf of themselves and all others similarly situated, **PLAINTIFFS,** **v.** **PUBLICIS GROUPE SA and MSLGROUP,** **DEFENDANTS.**	**Civ No. 11-CV-1279 (ALC) (AJP)**

DECLARATION OF PAUL J. NEALE IN SUPPORT OF PLAINTIFFS' REPLY IN SUPPORT OF RULE 72(a) OBJECTIONS TO MAGISTRATE JUDGE PECK'S FEBRUARY 8, 2012 DISCOVERY RULINGS

I, Paul J. Neale, declare as follows:

1. I am the Chief Executive Officer and a Managing Director of DOAR Litigation Consulting LLC and have been retained by Sanford Wittels and Heisler, LLP as a consultant and expert in the above-captioned matter.

2. I hold a Bachelor of Arts degree in criminal justice from Temple University.

3. I have advised lawyers and their clients on the management of information in litigation for over 20 years and am a nationally recognized expert on issues relating to the management and production of electronically stored information ("ESI").

4. I am a frequent author, lecturer and CLE instructor regarding the proper management of ESI and on the evolving state of the law and technology as they relate to ESI issues.

5. As a Managing Director at DOAR Litigation Consulting, I am routinely called upon to render expert advice and provide expert testimony on behalf of clients on discovery issues such as ESI preservation, spoliation, cost-shifting, reasonableness, inaccessibility determinations, ESI sanctions and the use of alternative technologies in the analysis and review of ESI.

6. I submit this declaration to clarify Plaintiffs' position and to address the misstatements and misrepresentations made in the declarations attached to Defendant MSL's brief, by two representatives of their vendor Recommind: Eric Seggebruch and Jan Puzicha.

7. Defendants, along with their experts at Recommind, obfuscate the flaws in the ESI protocol ("the protocol") adopted by Judge Peck, by focusing on the training of the Axcelerate system (i.e. Recommind's proprietary technology that is used in the ESI protocol), and by relying on the accuracy of other systems designed to conduct computer-assisted review, to support the accuracy of Recommind's Axcelerate system.

8. The protocol's primary flaw is that it does not include a scientifically supported method for validating the results of the Axcelerate system's predictive coding process as modified by the Defendants and accepted by Judge Peck. As it currently stands, the Plaintiffs, the Defendants and the Court will never know whether the Defendants' predictive coding process met any acceptable standard for the production of documents responsive to Plaintiffs' document requests.

9. As stated during the February 8, 2012 hearing and cited in Judge Peck's opinion, I am a proponent of the use of predictive coding, when it can be validated as reliable. However, the use of predictive coding or any other computer-assisted review approaches (including the use of keyword searching, which is also a type of computer-assisted review) should include a proper validation of the process against some pre-established measure within the context of the specific use of that approach.

10. The Defendants' and Judge Peck's reasoning that predictive coding is better than the alternatives, despite the lack of foundation, should not be a *de facto* validation of the Defendants' specific use of Recommind's Axcelerate system in the instant action. There must be some requirement to validate the efficacy of the process.

<u>Misstatements& Misrepresentations about the 2011 TREC Study</u>

11. Mr. Seggebruch's statements in his March 7, 2012 declaration (that were also echoed by Recommind in its widely distributed marketing material), which refer to Recommind's performance in the 2011 TREC study, are misleading and incomplete.

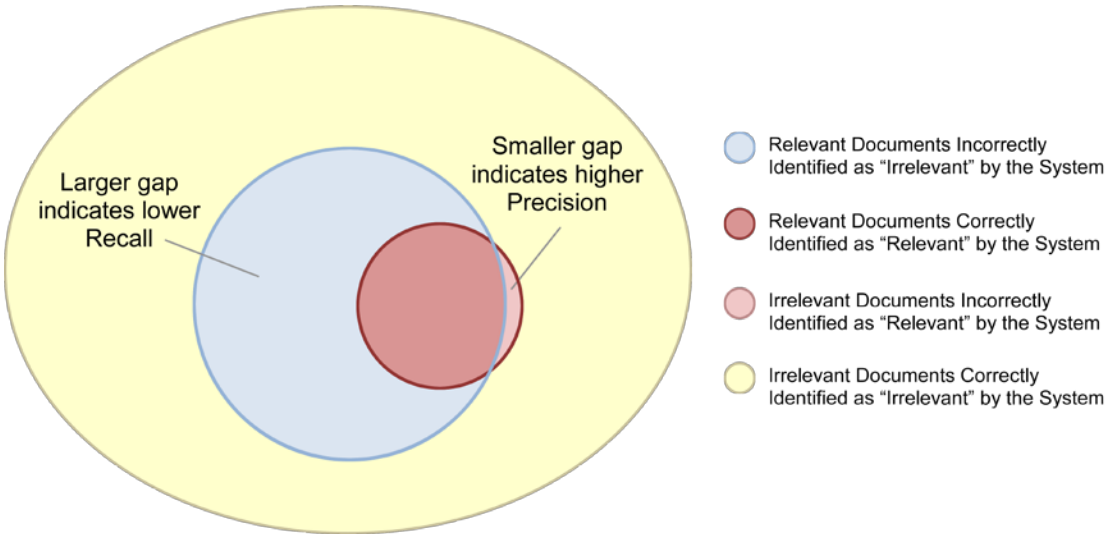
12. Mr. Seggebruch stated in paragraph 18 of his declaration: "In one category, Recommind achieved F_1 scores over 60%." This is not a very high score; one could reasonably infer from it that 40% of all responsive documents are likely to be missed by the Axcelerate system. More concerning, the score he reported referred to overall accuracy, but excluded "recall" – the proportion of responsive documents actually found.

13. In addition, Mr. Seggebruch failed to reveal to the Court in his declaration that the F_1 scores reported refer to Recommind's representation (not TREC's findings) of Recommind's "hypothetical F_1 scores." The hypothetical scores refer to an after-the-fact assessment of how the system would perform under the best possible circumstances, **not** how the system actually performed.

14. A draft version of TREC's report of the 2011 study (attached as Ex. 1) indicates that Recommind's **actual F_1 scores were significantly lower** than their hypothetical F_1 scores.[1] For example, for the same run in which Recommind received a 62.3% hypothetical F_1 score, their actual F_1 score is 24.7% and their recall was 25.8%. In other words, **over 74% of all responsive documents were missed**. (Ex. 1, at 13 Table 10.)

15. On average, Recommind's recall scores were approximately 35% in the 2011 TREC study.

16. Applying Recommind's recall scores from the 2011 TREC study to this case, Defendants would **fail to produce 65 out of every 100 of the relevant documents**. In other words, the system may incorrectly code the vast majority of relevant documents as "irrelevant." This is illustrated as follows:


Larger gap indicates lower Recall
Smaller gap indicates higher Precision
Relevant Documents Incorrectly Identified as "Irrelevant" by the System
Relevant Documents Correctly Identified as "Relevant" by the System
Irrelevant Documents Incorrectly Identified as "Relevant" by the System
Irrelevant Documents Correctly Identified as "Irrelevant" by the System

17. In my opinion, incorrectly identifying many relevant documents as "irrelevant" is an **unacceptable result**, even when you compare it to human review of documents as was done by Recommind and Defendant MSL in their brief, and Judge Peck in his written opinion.

[1] This document was not released by TREC and is admittedly preliminary and subject to change. Due to Recommind's selective usage of the results, however, it is, absent Recommind's own disclosure, the only evidence of Recommind's performance during the 2011 TREC study.

18. Mr. Seggebruch's dismissal of the TREC studies is contradicted by his company's marketing material. Recommind's material states, "In the final results stage where teams worked among themselves in a *real-world scenario*, Recommind had the best results (the highest accuracy) in all three topics, all by a wide margin." (Ex. 2 at 3 (emphasis added); *see also generally* Ex. 3.) Despite this publication, Mr. Seggebruch comfortably stated to the Court in paragraph 18 of his declaration, "I believe that the reliance on TREC is wholly misplaced because TREC is an academic exercise, rather than a real-world review."

19. Mr. Seggebruch's statements referring to Recommind's 2011 TREC results are further undermined by the fact that Recommind has been banned from future participation in TREC studies due to violating their agreement with TREC by publicizing its preliminary results as compared to other participants in 2011 and prior years.

Misstatements& Misrepresentations about the 2009 TREC Study

20. Defendant MSL improperly refers to the 2009 TREC study and to the article by Maura Grossman and Gordon Cormack that analyzes the results of that study as follows:[2]

    a. Recommind did not participate in the 2009 TREC study so the results in no way reflect their performance.

    b. The two systems that the Grossman & Cormack article determined to have been "conclusively superior" in the 2009 TREC study – H5 and the University of Waterloo – use technologies and methods that are distinctly different from Recommind' Axcelerate system that Judge Peck adopted.

21. It is a misrepresentation of fact for Defendant MSL to imply that the results of these two participants in the 2009 TREC study – a study in which Recommind did not participate – can be used to support the reliability of the protocol in this case. The notion that if one computer-assisted review system performs well, then all computer-assisted review systems must perform well is akin to Toyota using BMW's and Audi's safety tests to validate the safety of Toyota's vehicles. TREC 2009 and the Grossman & Cormack article do not evaluate Recommind's system, and thus cannot be used to support the potential performance of the Axcelerate program generally or in the context of this case.

[2] Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII Rich. J.L. & Tech. 11 (2011), http://jolt.richmond.edu/v17i3/article11.pdf. (Doc. 100.)

Misstatements & Misrepresentations about the "Quality Control" of the ESI Protocol

22. The protocol as currently adopted by Judge Peck does not include a reliable measurement of the accuracy of the protocol's system, and specifically fails to include a scientifically supportable measure of "recall," which is the metric that establishes what proportion of the responsive documents the system identified as "relevant."

23. Instead, the protocol takes a random sample of only 2,399 documents at the final stage of the system's "testing" in the section "Quality Control by Random Sample of Irrelevant Documents." This step in the predictive coding process is the final and only gauge as to whether Recommind's system is identifying as relevant an acceptable percentage of the responsive documents.

24. Using Defendant MSL's own numbers, a random sample of only 2,399 documents is not scientifically supportable. Brett Anders, counsel for Defendant MSL, stated at the January 4, 2012 hearing that his review of an initial random sample of documents indicated that the ultimate percentage of responsive documents would be 1.5% of the total population. (*See* Tr. Jan. 4, 2012 H'ring (Nurhussein Decl. Ex. A) at 46.) Based on this 1.5% number, one can predict that, of the approximately 2.5 million documents subject to the ESI protocol, 37,500 documents would be responsive to Plaintiffs' requests, and thus identified as "relevant." Conversely, the remaining 2,462,500 documents would be identified as "irrelevant" by the system.

25. The assessment of how many responsive documents (i.e. relevant documents) were missed by the system allows for the measurement of recall. The Court should note that **<u>the smaller the percentage of responsive documents in a given population, the larger the sample size required to measure recall.</u>**

26. A sample size of 2,399 documents randomly selected from 2,462,500 documents is not a statistically valid sample size that will allow the parties or the Court to determine how many responsive documents were missed by the Recommind system.

27. Mr. Puzicha, in paragraph 9 of his declaration, states that "a precise estimate of recall is irrelevant, however, as long as the estimation interval is within boundary of a standard accepted by the Court." The fact is, however, there has been no standard established by the Court. That is exactly what we are asking the Court to do.

28. In my opinion, **<u>a sample size of 16,555 documents</u>** during the "Quality Control by Random Sample of Irrelevant Documents" stage of the protocol **<u>is a statistically valid sample size</u>**, which is necessary given (1) the Defendants' estimation of a low yield percentage (1.5%) of relevant documents; and (2) preliminary evidence of Recommind's poor recall results in the 2011 TREC study

as discussed further above. Plaintiffs previously proposed this number as the sample size in its draft version of the protocol submitted to Judge Peck on January 3, 2012.

29. Furthermore, based on my over twenty years of experience advising clients on the review of document collections, I believe that the request to review 16,555documents is a small burden on Defendant MSL in the context of a collection containing 2.5 million documents.

I declare under the penalties of perjury that the foregoing is true and accurate to the best of my knowledge and belief.

Dated: March 19, 2012

Paul J. Neale

Exhibit 1

Overview of the TREC 2011 Legal Track
Notebook Draft 2011.10.24

Maura R. Grossman
Gordon V. Cormack
Bruce Hedin
Douglas W. Oard

Abstract

The TREC 2011 Legal Track consisted of one task: the *learning task*, which captured elements of the TREC 2010 learning and interactive tasks. Participants were required to rank the entire corpus of 670,000 documents by their estimate of the probability of relevance to each of 3 topics, and also to give a quantitative estimate of that probability. Participants were permitted to request up to 1,000 *relevance determinations* from a *Topic Authority* for each topic. Participants elected either to use only these relevance determinations in preparing *automatic* submissions, or to augment these determinations with their own manual review in preparing *technology assissted* submissions. We provide a brief overview of the task, and preliminary results as of October 24, 2011. More detailed results will be available to TREC participants during the conference at the web address `http://plg1.uwaterloo.ca/trec11-assess` .

1 Introduction

We are concerned with the document selection and review component of the *e-discovery* process, for which the objective is to identify as nearly as practicable all documents from a collection that are responsive to a *request for production* in civil litigation, while minimizing the number of unresponsive documents that are identified by the method.

The learning task models the scenario in which a senior attorney – the *Topic Authority* – is charged with interpreting the request for production, communicating that interpretation to a review team, and producing responsive documents to the requesting party. TREC participants play the role of the review team.

At the outset, the Topic Authority reads the request and prepares a set of coding guidelines. The request and the guidelines are given to participants, and an initial *kick-off call* allows interested participants to ask the Topic Authority how to interpret the request for production.

Over the course of several weeks, each participant is entitled to request feedback from the Topic Authority on a number of documents from the collection. This feedback consists of a simple binary *relevance determination:* Participants are informed whether the Topic Authority determines each document to be responsive or not. No other communication with the Topic Authority is permitted.

Teams from ten different groups participated in the Legal Track; the names of the teams, as well as the prefix used to label the team's results, are shown in table 1.

2 Document Collection

The document collection for the TREC 2011 Legal Track is identical to that used for TREC 2010. It was derived from the EDRM Enron Dataset, version 2, prepared by ZL Technologies in consultation with the Legal Track coordinators, and hosted by EDRM. The EDRM dataset consists of 1.3 million email messages captured by the Federal Energy Review Commission (FERC) from Enron, in the course of its investigation of Enron's collapse. ZL acquired the dataset from Lougheed Systems (formerly Aspen Systems) who captured

Participating Organization	Run Prefix
Beijing University of Posts and Telecommunications	pri
Helioid	HEL
Indian Statistical Institute	ISI
OpenText	ot
Recommind	rec
TCDI	TCD
University of Melbourne	mlb
University of South Florida	USF
University of Waterloo	UW
Ursinus College	URS

Table 1: Organizations participating in the TREC 2011 Legal Track.

and maintain the dataset on behalf of FERC. The EDRM dataset is available in two formats: EDRM XML and PST. The EDRM XML version contains a text rendering of each email message and attachment, as well as the original native format. The PST version contains the same messages, in a Microsoft proprietary format used by many commercial tools.

Both versions of the dataset approach 100GB in size, presenting an obstacle to participants. Furthermore, there are a large number of duplicate email messages in the dataset, that were captured more than once by Lougheed. For TREC, a list of 470,000 distinct messages were identified as canonical; all other messages duplicate one of the canonical messages. These messages contain about 200,000 attachment files; together these 470,000 messages plus 200,000 attachments form the 670,000 documents of the TREC 2010/2011 Legal Track collection. Text and native versions of these documents were made available to participants, along with a mapping from the EDRM XML and PST files to their canonical counterparts in the TREC collection.

3 Relevance Assessments

In order to measure the efficacy of TREC participant efforts in the two tasks, it is necessary to compare their results to a *gold standard* indicating whether or not each document in the collection is relevant to a particular discovery request. The learning task used three distinct discovery requests. Ideally, a gold standard would indicate the relevance of each document to each topic.

It is impractical to use human assessors to render these two million assessments. Instead, a sample of documents was identified for each topic, and assessors were asked to code only the documents in the sample as relevant or not. The evaluation was conducted in two phases: preliminary and final. At the time of writing, only the preliminary evaluation was complete.

3.1 Preliminary Evaluation

A total of 16,999 documents – about 5,600 per topic – were selected and asessed to form the preliminary gold standard. The documents were selected according to four criteria:

1. All documents that were identified by the coordinators, in the course of composing the topics before the start of the task, to be potentially relevant;
2. All documents submitted by any team for relevance determination;
3. All documents ranked among the 100 most probably relevant by any submission;
4. A uniform random sample of the remaining documents.

11,612 documents (the *100 stratum*) were selected according to one or more of the first three criteria; 5,387 documents (the *1000 stratum*) were sampled according to the fourth. All documents in the 100 stratum were assessed, regardless of whether or not a relevance determination had been previously rendered by the Topic Authority. Each document in the 1000 stratum was given to two assessors; that is, each sampled document was assessed twice.

The learning task assessments were rendered by four professional review companies, who volunteered their services. Three of the companies used a Web-based platform developed by the coordinators to view scanned documents and to record their relevance judgments. To avoid problems with local rendering software on each assessor's workstation, the assessors made their judgments based on pdf-formatted versions of the documents, as opposed to the original native format documents. The fourth review company downloaded the pdf documents and conducted the review on their own platform.

Assessors were provided with orientation and detailed guidelines created by a Topic Authority. The review platform included a "seek assistance" link which assessors were encouraged to use to request that the Topic Authority resolve any uncertainty that may have arisen. Assessors were instructed to make a relevance judgment of relevant (R), not relevant (N), or broken (B) for every document in their bins. The latter code reflects the fact that a small percentage of documents from the EDRM dataset are malformed and therefore cannot be assessed.

Once the preliminary assessments were complete, quality assurance was conducted by having the Topic Authority adjudicate conflicting assessments, which occurred in two cases:

1. For documents selected according to criterion 2 above, the Topic Authority's initial relevance determination and the assessor's relevance judgment differed; or

2. For documents selected according to criterion 4 above, i.e. the 1000 stratum, the two assessors' judgments differed.

The Topic Authority adjudicted all conflicting documents together, with no indication of which documents had been subject to a previous relevance determination, or what that determination had been.

The preliminary gold standard consists of

- The assessor's judgment, for documents without conflicting assessments; and,

- The Topic Authority's final judgment, for documents with conflicting assessments.

The preliminary gold standard, along with the toolkit used for the preliminary evaluation, may be found on the web: `http://durum0.uwaterloo.ca/trec/legal10-results` .

3.2 Final Evaluation

At the time of writing, additional assessments and quality assurance measures are being undertaken to improve the accuracy and reliability of the evaluation results. The set of documents identified by any submission as one of the 5,000 most probably relevant – the *5000 stratum* – has been sampled for assessment by the professional review companies. For topic 402, a second sample of the 1000 stratum is also being assessed. Additional redundant assessments are being conducted, and the toolkit is being enhanced to adjust the estimated evaluation measures to compensate for random errors in relevance assessments.

Final results will be available to participants at the TREC workshop in November and, at the same time, on the web: `http://plg1.uwaterloo.ca/trec11-assess` .

4 The Task

The learning task models the use of automated or semi-automated methods to guide review strategy for a multi-stage document review effort, organized as follows:

1. **Initial search and assessment.** The responding party analyzes the production request. Using ad hoc methods the team identifies a *seed set* of potentially responsive documents, and assesses each as responsive or not.

2. **Learning by example.** A learning method is used to rank the documents in the collection from most to least likely to be responsive to the production request, and to estimate this likelihood for each document. The input to the learning method consists of the seed set, the assessments for the seed set, and the unranked collection; the output is a ranked list consisting of the document identifier and a probability of responsiveness for each document in the collection.

   The two components of learning by example – ranking and estimation – may be accomplished by the same method or by different methods. Either may be automated or manual. For example, ranking may be done using an information retrieval method or by human review using a five-point scale. Estimation may be done in the course of ranking or, for example, by sampling and reviewing documents at representative ranks.

3. **Review process.** A review process may be conducted, with strategy guided by the ranked list. One possible strategy is to review documents in order, thus discovering as many responsive documents as possible for a given amount of effort. Another possible strategy is triage: to review only mid-ranked documents, deeming, without further review, the top-ranked ones to be responsive, and the bottom-ranked ones to be non-responsive.

   Review strategy may be guided not only by the order of the ranked list, as outlined above, but also by the estimated effectiveness of various alternatives. Consider the strategy of reviewing the top-ranked documents. Where should a *cut* be made so that documents above the cut are reviewed and documents below are not? For triage, where should the two necessary cuts be made?

Practically every review strategy decision boils down to the question,

> *Of this particular set of documents, how many are responsive and how many are not?*

This question itself can be reduced to,

> *What is the probability of each document in the set being relevant?*

Given an answer to the second question, the answer to the first is simply the sum of the probabilities. For this reason, participants in the learning track were required to provide an estimate of the probability of relevance for each document in the collection. Using these estimates the documents were ranked from most likely to least likely relevant. At each rank, the estimated number of relevant documents – the sum of the probabilities up to that rank – was computed, and used to estimate recall, precision and F_1.

4.1 Submission Phases

For each topic, teams were required to submit an *initial* set of probability estimates prior to requesting any relevance determinations from the Topic Authority. Thereafter, teams were required to submit *interim* sets of probability estimates in order to receive more than 100 and more than 300 relevance determinations. A team was allowed to request at most 1,000 relevance determinations per topic.

Each team was required to submit a *final* set of probability estimates once it had received all the relevance determinations requested by the team.

In a final *mopup* phase, all relevance determinations requested by all teams were distributed to all teams, who had the opportunity to submit a *mopup* set of probability estimates.

In this overview, we report results separately for the *initial*, *final* and *mopup* submissions. The run identifiers for the various phases may be distinguished by their final symbol: initial submissions end in “1”; final submissions end in “F”; and mopup submissions end in “M”.

4.2 Participation Categories

Participants were asked to declare each run to be *automatic* or *technology assisted.* Automatic runs were allowed to use manual query formulation, but human review of the document collection (other than that

Topic	Number of Responsive Documents
401	30,853
402	1,920
403	1,239

Table 2: Estimated number of responsive documents for each topic.

provided by TREC via responsiveness determinations) was not allowed. Technology assisted runs were allowed to avail themselves of any amount of human review. Participants were asked to state the number of documents reviewed, as well as the number of hours spent – both reviewing documents and configuring the system. The participation category of a run is specified by the penultimate character in its name: "A" for automatic; and "T" for technology assisted. For example, the run named gggxxxAF is a final run, automatic participation, by the group whose run prefix is ggg.

4.3 Topics

There were three topics: 401, 402 and 403.

- Topic 401 (Topic Authority: Kevin F. Brady, Connolly Bove Lodge & Hutz LLP.)
  All documents or communications that describe, discuss, refer to, report on, or relate to the design, development, operation, or marketing of enrononline, or any other online service offered, provided, or used by the Company (or any of its subsidiaries, predecessors, or successors-in-interest), for the purchase, sale, trading, or exchange of financial or other instruments or products, including but not limited to, derivative instruments, commodities, futures, and swaps.

- Topic 402 (Topic Authority: Brendan M. Schulman, Kramer Levin Naftalis & Frankel LLP.)
  All documents or communications that describe, discuss, refer to, report on, or relate to whether the purchase, sale, trading, or exchange of over-the-counter derivatives, or any other actual or contemplated financial instruments or products, is, was, would be, or will be legal or illegal, or permitted or prohibited, under any existing or proposed rule(s), regulation(s), law(s), standard(s), or other proscription(s), whether domestic or foreign.

- Topic 403 (Topic Authority: Robert Singleton, Squire, Sanders & Dempsey (US) LLP.)
  All documents or communications that describe, discuss, refer to, report on, or relate to the environmental impact of any activity or activities undertaken by the Company, including but not limited to, any measures taken to conform to, comply with, avoid, circumvent, or influence any existing or proposed rule(s), regulation(s), law(s), standard(s), or other proscription(s), such as those governing environmental emmissions, spills, pollution, noise, and/or animal habitats.

5 Preliminary Results

Based on the preliminary gold standard, we estimate the total number of responsive documents in the collection to be as shown in Table 2.

The following sections detail the results achieved by the various submissions. We note that not all teams undertook every topic, and not all teams submitted initial or mopup submissions. We note further that some recall estimates in the summary statistics exceed 1.000, due to random error in the preliminary evaluation.

5.1 Gain Curves

A *gain curve* shows the fraction of responsive documents (*recall*) as a function of the number of documents produced, when the documents are produced in order from most likely to least likely relevance. Thus each submission has an associated gain curve. Figure 1 shows the gain curves for the Topic 401 initial

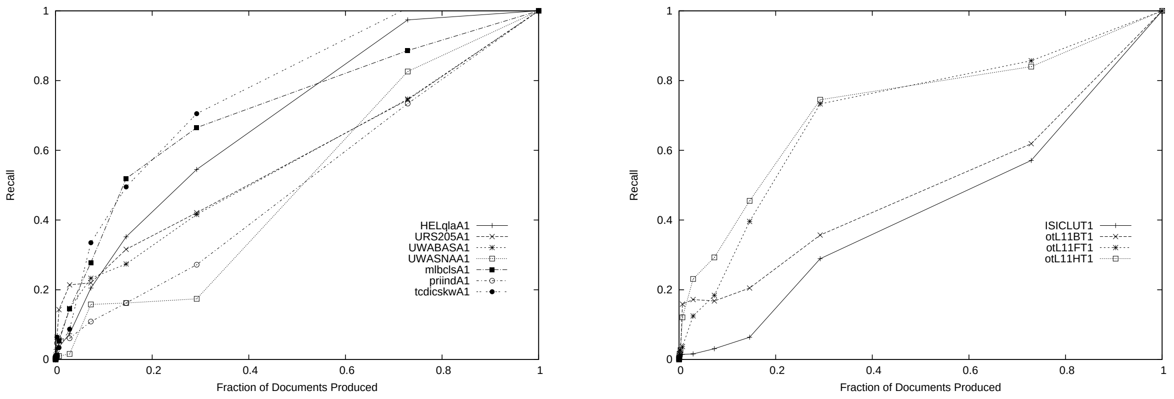


Figure 1: Gain curves for Topic 401 initial submissions (Left: automatic; Right: technology-assisted).

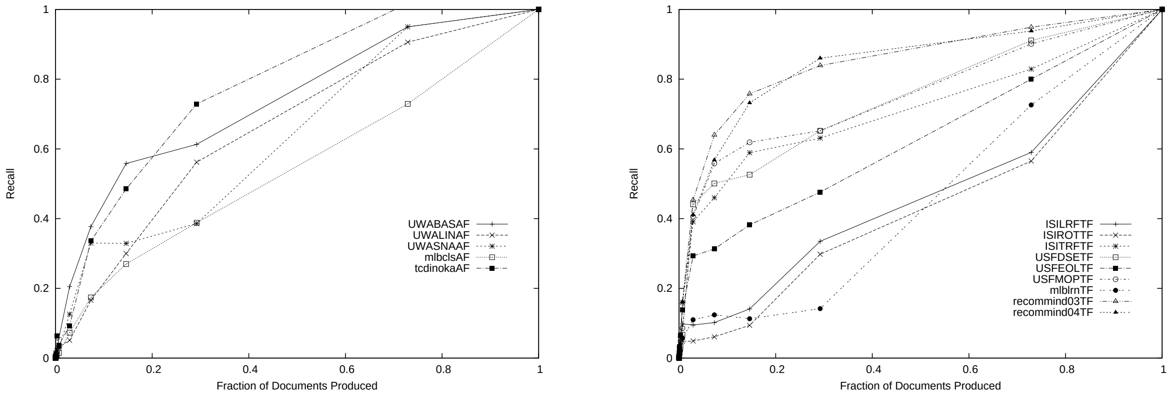


Figure 2: Gain curves for Topic 401 final submissions (Left: automatic; Right: technology-assisted).

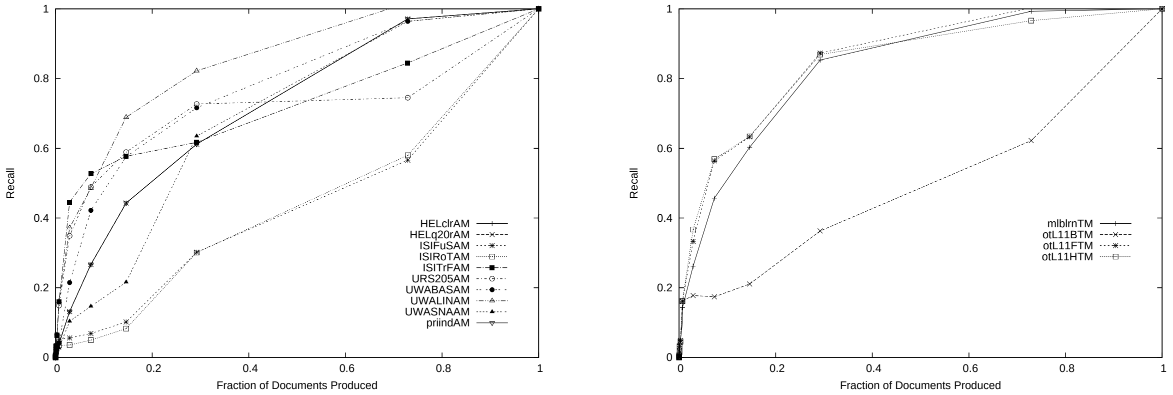


Figure 3: Gain curves for Topic 401 mopup submissions (Left: automatic; Right: technology-assisted).

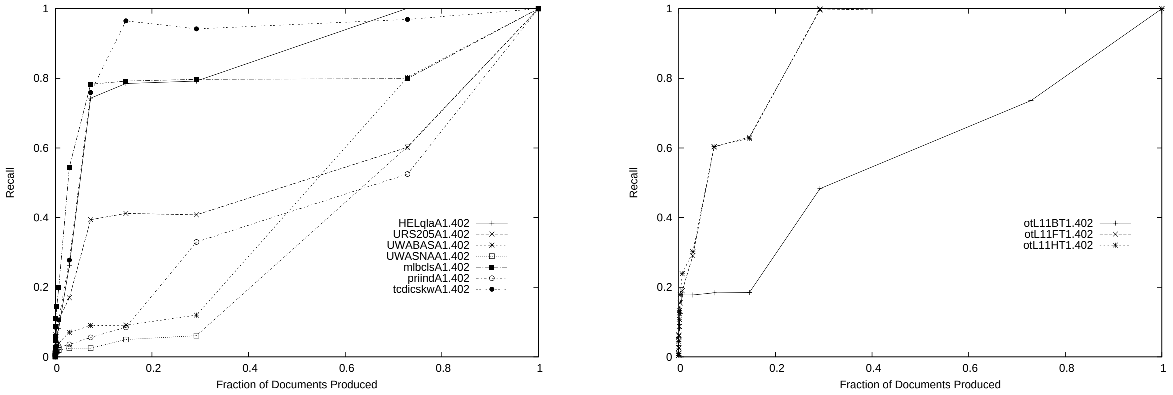


Figure 4: Gain curves for Topic 402 initial submissions (Left: automatic; Right: technology-assisted).

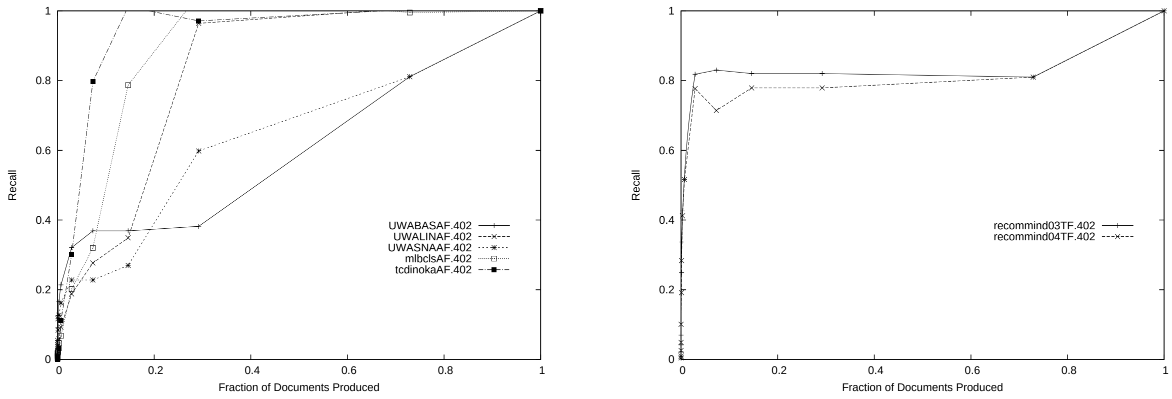


Figure 5: Gain curves for Topic 402 final submissions (Left: automatic; Right: technology-assisted).

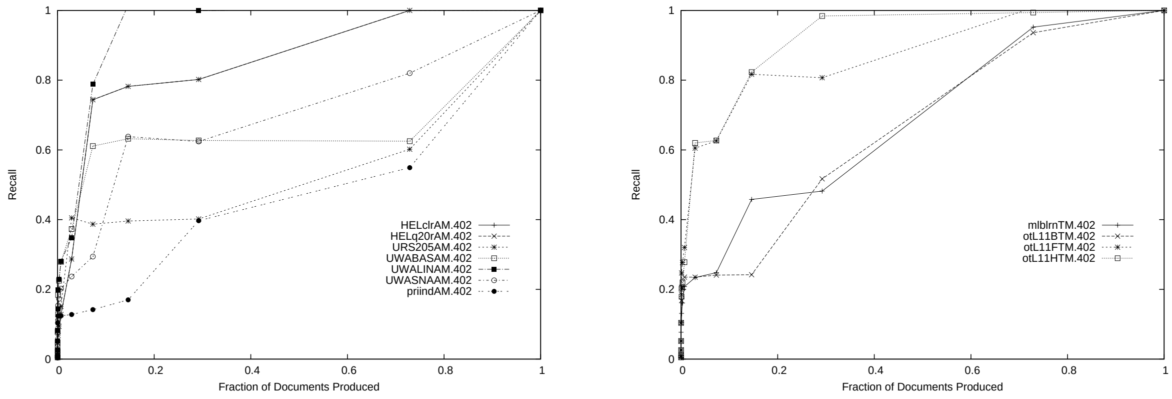


Figure 6: Gain curves for Topic 402 mopup submissions (Left: automatic; Right: technology-assisted).

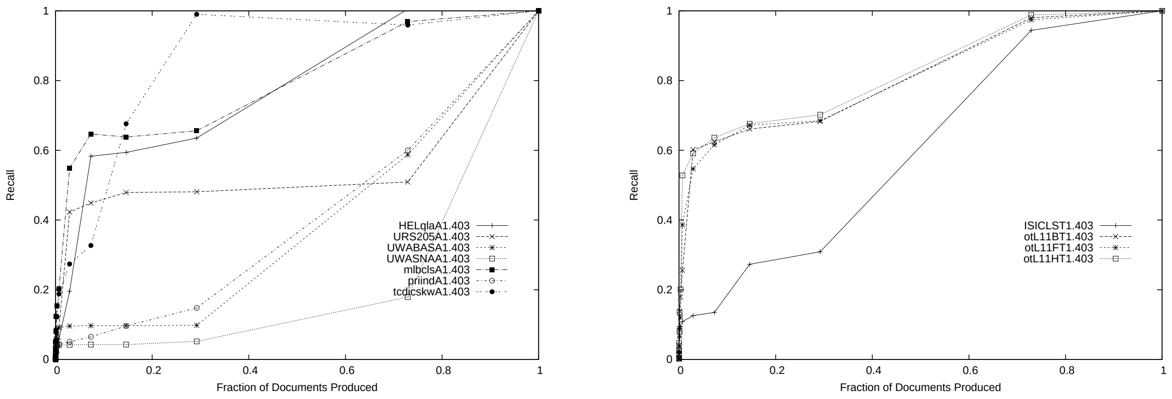


Figure 7: Gain curves for Topic 403 initial submissions (Left: automatic; Right: technology-assisted).

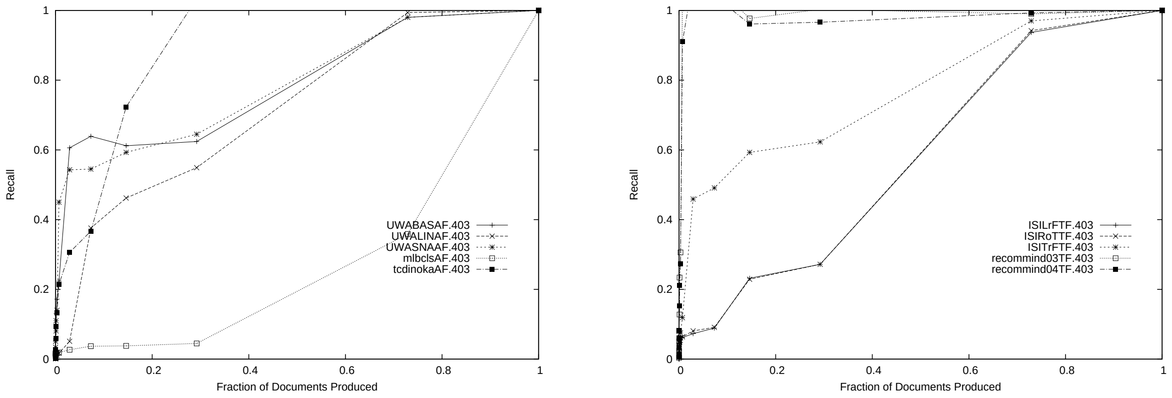


Figure 8: Gain curves for Topic 403 final submissions (Left: automatic; Right: technology-assisted).

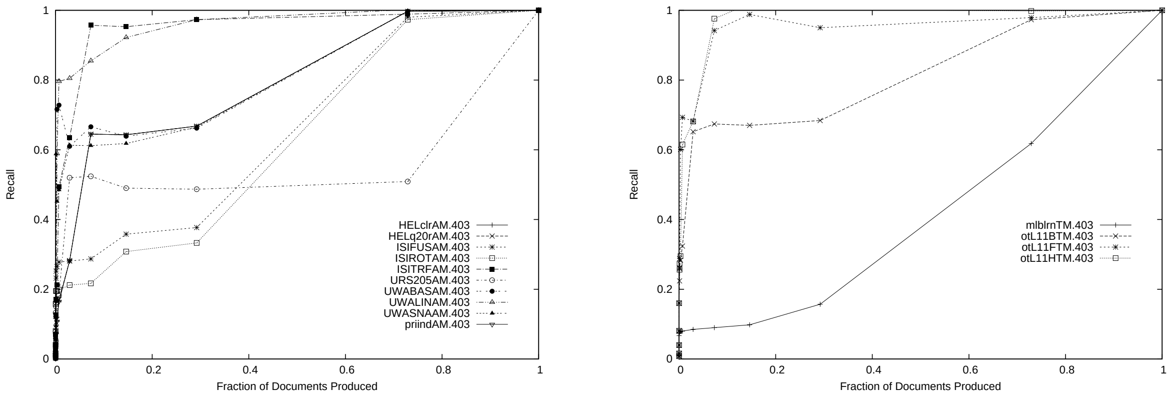


Figure 9: Gain curves for Topic 403 mopup submissions (Left: automatic; Right: technology-assisted).

submissions. The automatic submissions are shown on the left and the technology-assisted submissions on the right. Figures 2 and 3 show the corresponding curves for the final and mopup submissions.

Figures 4 through 6 show the corresponding information for Topic 402; Figures 7 through 9 for Topic 403.

5.2 Precision, Recall and F_1

Because participants submitted a probability estimate rather than a set of responsive documents, it is necessary to specify a *cutoff rank*, c, such that the c documents with the highest probability of relevance are deemed to be responsive, and the remainder are deemed to be non-responsive. Once the cutoff rank is chosen, recall, precision and F_1 may be calculated in the normal way.

For this set of evaluation measures we assume that the cutoff rank is to be chose so as to maximize F_1. Once the gold standard has been constructed, it is a simple matter to try all possible values of c and to choose the one that yields the maximum value of F_1. We call this value the *hypothetical* F_1, because it could be achieved, but only if the appropriate value of c were somehow determined without the aid of the gold standard.

Since c must be determined without knowledge of the gold standard, we must rely on the probability estimates contained in the submissions. As noted above, the probability estimates for the top-ranked c documents may be summed to yield an estimate of the number of responsive documents were the cut to be made at c. From this estimate we may easily derive estimates of recall, precision and F_1, and select the value of c that yields the largest F_1 estimate. We call this value the *actual* F_1, as it can be achieved using only the information contained in the submission.

Tables 3 through 11 show the hypothetical and actual F_1values, along with the recall and precision values at the cutoff rank that achieves the corresponding F_1 value. Separate tables are shown for initial, final and mopup runs, and for each topic.

6 Discussion

It is apparent from the gain curves that the best systems are able to identify the vast majority of responsive documents with a cutoff value that includes only a tiny fraction of the collection. For Topic 401, this level of recall is achieved with a cutoff of about 10%. For topics 402 and 403, the number is much lower.

	Hypothetical			Actual		
run	F_1	Recall	Precision	F_1	Recall	Precision
otL11HT1	0.372	0.276	0.572	0.302	0.197	0.646
URS205A1	0.343	0.223	0.745	0.168	0.092	0.987
otL11BT1	0.324	0.199	0.868	0.284	0.177	0.727
tcdicskwA1	0.279	0.436	0.205	0.115	1.005	0.061
mlbclsA1	0.259	0.423	0.187	0.095	0.971	0.050
otL11FT1	0.213	0.746	0.124	0.148	0.167	0.133
UWABASA1	0.198	0.150	0.291	0.089	0.692	0.048
HELqlaA1	0.183	0.349	0.124	0.086	1.000	0.045
priindA1	0.127	0.070	0.724	0.111	0.061	0.600
UWASNAA1	0.122	0.168	0.096	0.091	0.689	0.049
ISICLUT1	0.088	0.990	0.046	0.000	0.000	0.875

Table 3: Topic 401 initial submission results.

	Hypothetical			Actual		
run	F_1	Recall	Precision	F_1	Recall	Precision
recommind03TF	0.588	0.585	0.591	0.476	0.319	0.939
recommind04TF	0.581	0.435	0.874	0.510	0.350	0.937
USFDSETF	0.560	0.486	0.660	0.128	0.077	0.370
USFMOPTF	0.543	0.481	0.623	0.176	0.113	0.406
ISITRFTF	0.505	0.433	0.606	0.000	0.000	0.875
USFEOLTF	0.426	0.322	0.631	0.376	0.242	0.847
UWABASAF	0.339	0.510	0.254	0.109	0.983	0.058
tcdinokaAF	0.286	0.439	0.212	0.088	0.995	0.046
UWASNAAF	0.253	0.331	0.205	0.109	0.983	0.058
mlblrnTF	0.179	0.111	0.457	0.040	0.021	0.322
ISILRFTF	0.176	0.103	0.602	0.000	0.000	0.125
UWALINAF	0.160	0.273	0.114	0.115	0.831	0.062
mlbclsAF	0.143	0.161	0.129	0.020	0.011	0.173
ISIROTTF	0.122	0.067	0.657	0.000	0.000	0.875

Table 4: Topic 401 final submision results.

	Hypothetical			Actual		
run	F_1	Recall	Precision	F_1	Recall	Precision
ISITrFAM	0.578	0.456	0.788	0.001	0.000	1.000
otL11HTM	0.505	0.397	0.694	0.332	0.200	0.997
otL11FTM	0.484	0.368	0.708	0.084	0.044	0.834
UWALINAM	0.458	0.401	0.533	0.272	0.796	0.164
URS205AM	0.446	0.409	0.491	0.427	0.378	0.492
mlblrnTM	0.377	0.359	0.398	0.376	0.355	0.398
UWABASAM	0.373	0.422	0.335	0.109	0.995	0.058
otL11BTM	0.337	0.208	0.893	0.318	0.193	0.896
HELq20rAM	0.246	0.451	0.169	0.142	0.950	0.077
HELclrAM	0.246	0.451	0.169	0.142	0.950	0.077
UWASNAAM	0.177	0.699	0.101	0.109	0.995	0.058
ISIFuSAM	0.141	0.076	0.992	0.001	0.000	1.000
priindAM	0.095	0.148	0.070	0.062	0.034	0.400
ISIRoTAM	0.088	0.989	0.046	0.001	0.000	1.000

Table 5: Topic 401 mopup submission results.

	Hypothetical			Actual		
run	F_1	Recall	Precision	F_1	Recall	Precision
otL11HT1	0.185	0.131	0.314	0.065	0.289	0.036
otL11FT1	0.168	0.130	0.238	0.047	0.308	0.025
mlbclsA1	0.156	0.139	0.177	0.006	0.992	0.003
otL11BT1	0.130	0.091	0.232	0.120	0.138	0.106
URS205A1	0.086	0.057	0.176	0.036	0.139	0.021
tcdicskwA1	0.078	0.736	0.041	0.007	0.970	0.004
HELqlaA1	0.078	0.052	0.155	0.006	1.000	0.003
UWABASA1	0.034	0.035	0.034	0.005	0.538	0.002
UWASNAA1	0.025	0.015	0.092	0.003	0.346	0.001
priindA1	0.023	0.015	0.053	0.022	0.026	0.019

Table 6: Topic 402 initial submission results.

	Hypothetical			Actual		
run	F_1	Recall	Precision	F_1	Recall	Precision
recommind03TF	0.588	0.423	0.960	0.477	0.384	0.629
recommind04TF	0.455	0.440	0.470	0.395	0.387	0.404
UWABASAF	0.172	0.156	0.191	0.007	1.005	0.004
UWASNAAF	0.154	0.117	0.225	0.007	1.004	0.004
tcdinokaAF	0.087	0.776	0.046	0.006	0.999	0.003
UWALINAF	0.059	0.067	0.053	0.020	0.752	0.010
mlbclsAF	0.046	0.047	0.046	0.037	0.186	0.020

Table 7: Topic 402 final submission results.

	Hypothetical			Actual		
run	F_1	Recall	Precision	F_1	Recall	Precision
otL11FTM	0.345	0.234	0.655	0.224	0.126	1.000
otL11HTM	0.289	0.192	0.581	0.286	0.190	0.580
otL11BTM	0.273	0.166	0.781	0.209	0.199	0.220
UWALINAM	0.262	0.201	0.377	0.091	0.746	0.048
UWABASAM	0.250	0.199	0.337	0.007	1.007	0.004
priindAM	0.221	0.124	1.000	0.082	0.124	0.061
UWASNAAM	0.209	0.162	0.293	0.007	1.007	0.004
mlblrnTM	0.175	0.142	0.226	0.123	0.066	0.858
URS205AM	0.158	0.100	0.374	0.068	0.395	0.037
HELclrAM	0.111	0.073	0.226	0.010	1.001	0.005
HELq20rAM	0.110	0.074	0.213	0.010	1.002	0.005

Table 8: Topic 402 mopup submission results.

		Hypothetical			Actual	
run	F_1	Recall	Precision	F_1	Recall	Precision
otL11HT1	0.335	0.580	0.235	0.106	0.539	0.059
otL11FT1	0.234	0.339	0.179	0.078	0.494	0.042
otL11BT1	0.188	0.552	0.114	0.104	0.587	0.057
mlbclsA1	0.149	0.434	0.090	0.004	0.986	0.002
ISICLST1	0.110	0.078	0.183	0.008	0.004	0.625
tcdicskwA1	0.098	0.149	0.073	0.005	0.958	0.002
URS205A1	0.083	0.352	0.047	0.033	0.105	0.019
UWABASA1	0.066	0.057	0.080	0.001	0.128	0.000
priindA1	0.063	0.036	0.234	0.019	0.044	0.012
UWASNAA1	0.035	0.040	0.031	0.001	0.122	0.000
HELqlaA1	0.032	0.578	0.016	0.004	1.000	0.002

Table 9: Topic 403 initial submission results.

		Hypothetical			Actual	
run	F_1	Recall	Precision	F_1	Recall	Precision
recommind03TF	0.623	1.332	0.407	0.247	0.258	0.237
recommind04TF	0.574	0.903	0.421	0.065	0.996	0.033
UWASNAAF	0.241	0.491	0.160	0.005	0.984	0.002
UWABASAF	0.146	0.127	0.171	0.005	0.984	0.002
ISITrFTF	0.139	0.443	0.083	0.008	0.004	0.625
tcdinokaAF	0.107	0.166	0.079	0.004	0.996	0.002
ISIRoTTF	0.090	0.048	0.787	0.008	0.004	0.625
ISILrFTF	0.046	0.032	0.078	0.000	0.000	0.000
UWALINAF	0.035	0.322	0.019	0.009	0.512	0.005
mlbclsAF	0.031	0.017	0.221	0.002	0.367	0.001

Table 10: Topic 403 final submission results.

		Hypothetical			Actual	
run	F_1	Recall	Precision	F_1	Recall	Precision
UWASNAAM	0.720	0.702	0.739	0.005	0.997	0.002
otL11FTM	0.612	0.664	0.567	0.345	0.215	0.875
UWABASAM	0.578	0.827	0.444	0.005	0.998	0.002
UWALINAM	0.467	0.660	0.362	0.048	0.820	0.025
otL11HTM	0.376	0.249	0.759	0.310	0.270	0.363
ISIFUSAM	0.340	0.228	0.670	0.013	0.006	1.000
otL11BTM	0.337	0.210	0.850	0.228	0.281	0.192
priindAM	0.325	0.195	0.984	0.123	0.195	0.090
ISIROTAM	0.325	0.195	0.984	0.013	0.006	1.000
ISITRFAM	0.249	0.508	0.165	0.013	0.006	1.000
URS205AM	0.239	0.145	0.672	0.043	0.521	0.022
mlblrnTM	0.134	0.073	0.734	0.055	0.028	0.795
HELq20rAM	0.093	0.097	0.090	0.007	1.004	0.003
HELclrAM	0.090	0.090	0.089	0.007	1.004	0.003

Table 11: Topic 403 mopup submission results.

The evaluation of recall, precision and F_1 shows overwhelmingly that the submitted probability estimates are very poor. If the probability estimates were accurate, they would yield an actual F_1 value close to the hypothetical F_1. Estimation is no mere academic exercise. A fundamental question in e-discovery is: when have I done enough? Estimation is essential to answering this question. The coordinators hope that these results will provide impetus to disover better estimation techniques.

As we have noted, these results are preliminary. It may be that errors in the gold standard cause system performances to be underestimated or (perhaps less likely) to be overestimated. The results contained here will be superceded by final results in which we will seek to mitigate these errors.

7 Conclusion

This is the sixth year of the TREC Legal Track, and our third year of building test collections based on Enron email [1, 5, 4, 3, 2]. Relevance judgments are now available for 14 topical production requests in addition to the (2010) review for privilege. The Legal Track will continue in 2012, when we anticipate having a new collection available for use by participants. We look forward to discussing what we have learned this year and the opportunities for 2012 when we meet this November in Gaithersburg.

References

[1] J. Baron, D. Lewis, and D. Oard. Trec 2006 legal track overview. In *Proc. 15th Text REtrieval Conference*, 2006.

[2] G. Cormack, M. Grossman, B. Hedin, and D. Oard. Overview of the trec 2010 legal track. In *Proc. 19th Text REtrieval Conference*, 2010. to appear.

[3] B. Hedin, D. Oard, S. Tomlinson, and J. Baron. Overview of the trec 2009 legal track. In *Proc. 18th Text REtrieval Conference*, 2009.

[4] D. Oard, B. Hedin, S. Tomlinson, and J. Baron. Overview of the trec 2008 legal track. In *Proc. 17th Text REtrieval Conference*, 2008.

[5] S. Tomlinson, D. Oard, J. Baron, and P. Thompson. Overview of the trec 2007 legal track. In *Proc. 16th Text REtrieval Conference*, 2008.

Exhibit 2

2011 TREC Legal Track FAQs

Table of Contents

What is TREC?

TREC is a conference co-sponsored by the by the United States National Institute of Standards and Technology (NIST) and Department of Defense. TREC stands for "Text REtrieval Conference" and has been run by NIST for many years to provide an environment for competitive measurement of systems and collaboration around various kinds of information retrieval tasks. For the last five years, TREC has run a Legal Track to promote adoption of technology to help improve the efficiency of the eDiscovery process. For more information go to: http://trec-legal.umiacs.umd.edu/

What was Tested in the 2011 TREC Legal Track Competition?

The 2011 Legal Track competition measured the performance of systems in identifying responsive documents in three different topics in the form of requests for production which were labeled 401, 402, and 403 and representative of typical document requests. The topics were designed to be both well-suited *and* ill-suited to technological assistance. The test was composed of two parts: the first part simulated a real review in that each team worked stand-alone to code all documents for responsiveness and submit these coding calls to TREC (called the "final results" stage); the second part, which did not mirror a typical review, enabled all teams to leverage the results of other teams' coding decisions to provide a baseline for further academic research (called the "mop-up" stage). The accuracy of each participating system was measured using F1 scores (an average measure of accuracy—see "What is an F1 Score" below). Results were assessed by professional review companies.

The 2011 TREC Legal Track competition was the most complete and competitive ever. A record number of teams signed up to participate in TREC 2011, submitting more than 45 different runs per topic. The top performing system from both the 2010 and 2009 TREC competitions competed in TREC 2011.

Recommind competed in the 2011 TREC Legal Track competition for the first time ever.

What were the Results of the 2011 TREC Legal Track Competition?

Recommind dominated TREC 2011. Recommind had the best results in both the final and mop-up stages of TREC 2011 by a wide margin. In the final results stage where teams worked among themselves in a real-world scenario, Recommind had the best results (the highest accuracy) in all three topics, all by a wide margin.

Recommind easily beat the top performer from the 2010 and 2009 competition. Recommind's Axcelerate system easily bested the top-performing system from the 2010 and 2009 TREC Legal Track competitions, who came in third in the 2011 TREC competition. In fact, the results were not close, with

Recommind's Axcelerate system proving to be 10 *times* more efficient than the next-best system in the 2011 competition—who themselves finished ahead of the top-performing system from 2010 and 2009. In fact, the Axcelerate system's efficiency superiority measured as high as 50x over competing systems in the TREC 2011 study.

Does Accuracy Matter?

Yes! **The TREC 2011 results showed that even small improvements in accuracy generate huge benefits in efficiency**. This is due to the fact that Axcelerate's greater accuracy during the seed set stage (commonly referred to as ECA) – where relevant documents are identified using Axcelerate's powerful analytics capabilities – is magnified many times over during the predictive coding process. The superior accuracy of the Axcelerate system thus resulted in enormous efficiency gains over all other competitors: across all three topics, Recommind's Axcelerate system was 10 *times* more efficient (9.86x, to be exact) than the next-best system. That means that in order to find the same number of responsive documents, using the next-best system would have required reviewing 10 *times* more documents. And remember that the 2011 runner-up system finished ahead of the system which won the TREC 2010 and 2009 competitions.

Does Efficiency Really Matter?

Absolutely – it is the key to document review**. On average, a legal team using Recommind's Axcelerate system would need to review *90% fewer* documents** to find the same number of responsive documents when using the next-best system – other systems are even worse. That translates directly into **90% lower document review costs**. The only way a legal team using another system could compete is by charging far more for review or drastically lowering the quality of legal services provided so that the review or investigation is less accurate. That means many privileged documents are being produced, the risk of spoliation sanctions skyrockets, and documents important to effective case planning are not being found.

Where Does the Recommind Advantage Come From?

The inherent advantage with using Recommind's Axcelerate system comes from its advanced PLSA technology combined with other machine-learning techniques for effective and accurate ECA, as well as Axcelerate's patented, lawyer-driven Predictive Coding workflow and superior predictive coding technology. No other vendor can replicate these factors.

The term "Predictive Coding" is being used in the market to mean many things; but as the 2011 TREC competition clearly showed, only Recommind provides the technology and patented process to provide first-class results.

Another Vendor Says They Didn't Compete This Year, but They're Almost as Good as Recommind…

As the above results clearly show, no one else is even close. All the significant vendors have competed in TREC from 2009-2011. The top-performing team from 2010 and 2009 also participated in and finished TREC 2011. That team finished third in the 2011 competition.

Recommind's Axcelerate system is 10 times more efficient than the next-best competitive system—which means that **Axcelerate is 10 times more efficient than any significant vendor in the market**.

What Is an F1 Score?

An F1 score is the harmonic mean of precision and recall scores. The F1 is an average measure of accuracy, given the two measures of accuracy, precision and recall.

What Are Precision and Recall?

Precision and recall are different measures of accuracy.

Precision is a percentage measure of how many items placed in a bucket (or a category, like "responsive"), actually belong there. Mathematically, this is:

$$\frac{\text{\# of actually responsive documents identified as responsive}}{\text{Total \# of documents identified as responsive}}$$

Recall is a percentage measure of how many items that should have gone in the bucket, were actually placed there. Mathematically, recall is:

$$\frac{\text{\# of actually responsive documents identified as responsive}}{\text{Total \# of responsive documents in the entire document collection}}$$

So recall can be thought of as a measure of completeness and precision as a measure of being right.

Exhibit 3

TREC 2011 Legal Track

2011 TREC Legal Track Most Competitive Ever

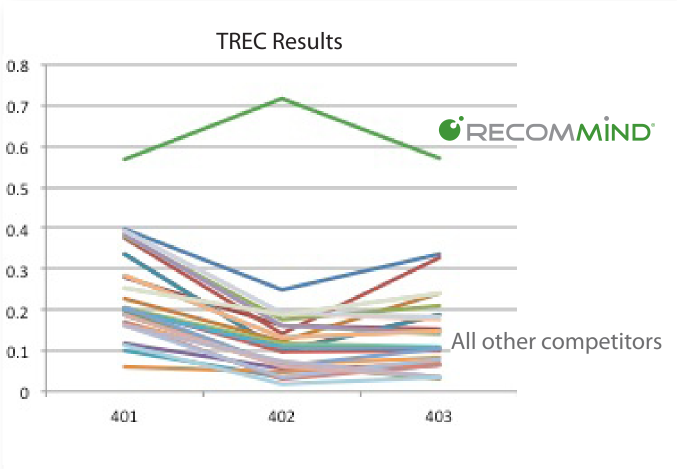
A record number of teams entered the 2011 TREC competition, submitting more than 45 different runs per topic. Participating teams included market-leading vendors as well as academic teams from universities all over the world. The 2011 finishers included the top-performing teams from both the 2010 and 2009 TREC study.

The evaluation was comprised of identifying and coding documents responsive to three different topics (labeled 401, 402, and 403) which were chosen for being representative of typical document requests.

TREC Results Confirm Recommind's Axcelerate® is Dramatically More Accurate & Efficient

In the 2011 TREC Legal Track, Recommind's Axcelerate system was the landslide winner across the board and in all categories tested. Participant results (as measured by F1 scores) showed a significant accuracy advantage for Axcelerate versus all other systems. Small accuracy advantages generated large efficiency gains. This led to Recommind being *10 times* more efficient than the 2nd place participant.

How Is Accuracy Measured?



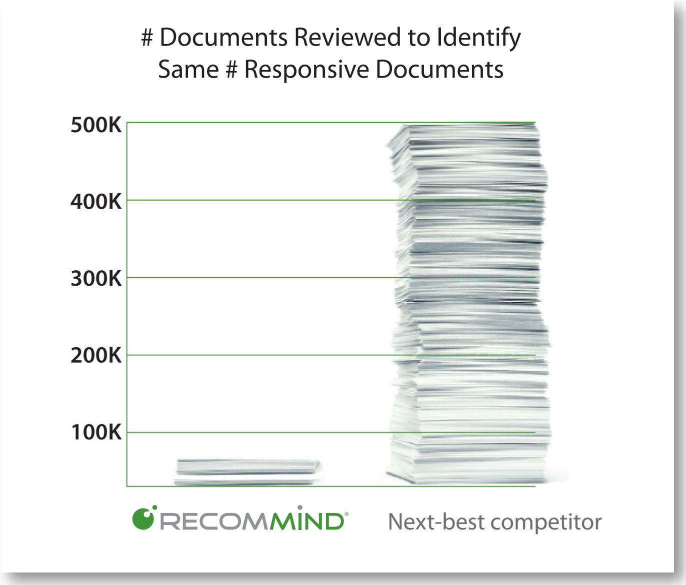
In TREC evaluations, accuracy is measured using two basic elements: precision and recall. *Precision* is a percentage measure of how many items placed in a category, like "responsive", actually belong there. *Recall* is a percentage measure of how many items that should have gone in the category, were actually placed there. The *F1 score* mentioned above is an average (the harmonic mean) of these two measures.

Not All Technology Is Created Equal

Recommind's patented PLSA algorithm combined with the patented Predictive Coding process and numerous other technologies integrated into the Axcelerate system provide a unique capability for performing highly effective investigations and document review. As the 2011 TREC results clearly showed, Recommind's Axcelerate system is dramatically more effective and efficient than any other product on the market.

Recommind's Clear Accuracy Advantages Lead to Enormous Efficiency Gains

The Axcelerate system's superior accuracy generated an enormous efficiency advantage. As the TREC 2011 results showed, a review or investigation using an Axcelerate system is on average 10x (9.86x, to be exact) more efficient than the next-best system. That means that in order to identify the same number of responsive documents, *10 times* more documents would have to be reviewed using the next-best system. Over other competitors, the efficiency gains are even higher – they can be as high as 50x.



What is TREC?

TREC—the "Text Retrieval Conference"— is co-sponsored by the National Institute of Standards and Technology (NIST) to provide a venue for scientifically evaluating information retrieval and coding systems. For the last 5 years, TREC has run a Legal Track to promote technology improvements in the legal industry. For more information go to: http://trec-legal.umiacs.umd.edu/

About Recommind

Recommind is the leader in predictive information management, delivering search-powered business applications that transform the way organizations conduct enterprise search, information governance and eDiscovery.

For more information please contact:

Recommind, Inc.
San Francisco, CA
www.recommind.com

sales@recommind.com
US Sales: +1 888 415 7899
EU Sales: +44 207 002 7735

RECOMMIND®
Out Predict. Out Perform.