Overview of the TREC 2007 Legal Track

Stephen Tomlinson, stomlins@opentext.com
Open Text Corporation, Ottawa, Ontario, Canada

Douglas W. Oard, oard@umd.edu
College of Information Studies and Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742, USA

Jason R. Baron, jason.baron@nara.gov
National Archives and Records Administration
Office of the General Counsel, Suite 3110, College Park, MD 20740, USA

Paul Thompson, Paul.Thompson@dartmouth.edu
Institute for Security Technology Studies
Dartmouth College, Hanover, NH 03755, USA

Abstract

TREC 2007 was the second year of the Legal Track, which focuses on evaluation of search technology for discovery of electronically stored information in litigation and regulatory settings. The track included three tasks: Ad Hoc (i.e., single-pass automatic) search, Relevance Feedback (two-pass search in a controlled setting with some relevant and nonrelevant documents manually marked after the first pass) and Interactive (in which real users could iteratively refine their queries and/or engage in multi-pass relevance feedback). This paper describes the design of the three tasks and analyzes the results.

1 Introduction

The use of information retrieval techniques in law has traditionally focused on providing access to legislation, regulations, and judicial decisions. Searching business records for information pertinent to a case (or "discovery") has also been important, but searching records in electronic form was until recently the exception rather than the norm. The goal of the Legal Track at the Text Retrieval Conference (TREC) is to assess the ability of information retrieval technology to meet the needs of the legal community for tools to help with retrieval of business records, an issue of increasing importance given the vast amount of information stored in electronic form to which access is increasingly desired in the context of current litigation. Ideally, the results of a study of how well comparative search methodologies perform when tasked to execute types of queries that arise in real litigation will serve to better educate the legal community on the feasibility of automated retrieval as well as its limitations. The TREC Legal Track was held for the first time in 2006, when 6 research teams participated in an ad hoc retrieval task. This year, 13 research teams participated in the track, which consisted of three tasks: 1) Ad Hoc, 2) Interactive, and 3) Relevance Feedback.

The results of the Legal Track are especially timely and important given recent changes in the U.S. Federal Rules of Civil Procedure that went into effect on December 1, 2006. The amended rules introduce a new category of evidence, namely, "Electronically Stored Information" (ESI) in "any medium," intended to stand on an equal footing with existing rules covering the production of "documents." Against the backdrop of the Federal Rules changes, the status quo in the legal profession, even in large and complex litigation, is continued reliance on free-text Boolean searching to satisfy document (and now ESI) production demands [6]. An important aspect of e-discovery and thus of the TREC Legal Track is an emphasis on recall over precision. In light of the fact that a large percentage of requests for production of documents (and now ESI) routinely state that "all" such evidence is to be produced, it becomes incumbent on responding parties to attempt to maximize the number of responsive documents found as the result of a search.

The key goal of the TREC Legal Track is to apply objective benchmark criteria for comparing search technologies, using topics that approximate how real lawyers would go about propounding discovery in civil litigation, and a large, representative (unstructured and heterogeneous) document collection. Given the reality of the use of Boolean search in present day litigation, comparing the efficacy of Boolean search using negotiated queries with alternative methods is of considerable interest. The Legal Track has shown that alternative methods do identify many relevant documents that were missed by a reference implementation of a Boolean search, though no single alternative method has yet been shown to consistently outperform Boolean search without increasing the number of documents to review.

The remainder of this paper is organized as follows. Section 2 describes the Ad Hoc task, Section 3 describes the Interactive and Relevance Feedback tasks, Section 4 lists the individual topic results, Section 5 summarizes the workshop discussions and analysis conducted after the conference, and Section 6 concludes the paper.

2 Ad Hoc Task

In the Ad Hoc task, the participants were given requests to produce documents, herein called "topics", and a set of documents to search. The following sections provide more details, but an overview of the differences from the previous year is as follows:

- At the time of topic release, the B value (the number of documents matching the final negotiated Boolean query) was provided for each topic in 2007, along with an alphabetical list (by document-id) of the documents matching the Boolean query (the "refL07B" run) for optional use by participants.

- A new evaluation measure, Estimated Recall@B, where B is the number of documents matching the Boolean query, was established as the principal measure for the track (although other measures are also reported). The legal community is interested in knowing whether additional relevant documents (those missed by a Boolean query) can be found for the same number of retrieved documents.

- A new sampling method (herein called "L07") was used to produce estimates of the main measure for each topic for all submitted runs. All runs submitted to the Ad Hoc task were pooled this year, and all pooled runs were treated equally by the sampling procedure.

- The new topics were vetted to ensure that the B value for any topic was in the 100 to 25,000 range. (In 2006, B ranged from 1 to 128,195.)

- Participating teams were allowed to submit up to 25,000 documents for each topic (up from 5,000 in 2006).

- To facilitate cross-site comparisons, a "standard condition" run which just used the (typically one-sentence) request text field was requested from all groups. Additional runs which used other topic fields were also welcome, and encouraged.

- Three different Boolean queries were provided for each topic (defendant, plaintiff and final). In 2006, the plaintiff and final queries had (usually) been the same.

2.1 Document Collection

The 2007 Legal Track used the same collection as the 2006 Legal Track, the IIT Complex Document Information Processing (CDIP) Test Collection, version 1.0 (referred to here as "IIT CDIP 1.0") which is based on documents released under the tobacco "Master Settlement Agreement" (MSA). The MSA settled a range of lawsuits by the Attorneys General of several US states against seven US tobacco organizations (five tobacco companies and two research institutes). One part of this agreement required those organizations to make public all documents produced in discovery proceedings in the lawsuits by the states, as well as all documents produced in a number of other smoking and health-related lawsuits. Notable among the provisions is that the organizations were required to provide to the National Association of Attorneys General (NAAG) a copy of metadata and the scanned documents from the websites, and are forbidden from objecting to any subsequent distribution of this material.

The University of California San Francisco (UCSF) Library, with support from the American Legacy Foundation, has created a permanent repository, the Legacy Tobacco Documents Library (LTDL), for tobacco documents [10]. The IIT CDIP 1.0 collection is based on a snapshot, generated between November 2005 and January 2006, of the MSA subcollection of the LTDL. The snapshot consisted of 1.5 TB of scanned document images, as well as metadata records and Optical Character Recognition (OCR) produced from the images by UCSF. The IIT CDIP project subsequently reformatted the metadata and OCR, combined the metadata with a slightly different version obtained from UCSF in July 2005, and discarded some documents with formatting problems, to produce the IIT CDIP 1.0 collection [8]. The IIT CDIP 1.0 collection consists of 6,910,192 document records in the form of XML elements.

IIT CDIP 1.0 has had strengths and weaknesses as a collection for the Legal Track. Among the strengths are the wide range of document genres (including letters, memos, budgets, reports, agendas, minutes, plans, transcripts, scientific articles, and email) and the large number of documents. Among the weaknesses are that the documents themselves were released as a result of tobacco-related discovery requests, and thus may exhibit a skewed topic distribution when compared with the larger collections from which they were initially selected. See the 2006 TREC Legal Track overview paper for additional details about the IIT CDIP 1.0 collection [3].

2.2 Topics

Topic development in 2007 continued to be modeled on U.S. civil discovery practice. In the litigation context, a "complaint" is filed in court, outlining the theory of the case, including factual assertions and causes of action representing the legal theories of the case. In a regulatory context, often formal letters of inquiry serve a similar purpose by outlining the scope of the proposed investigation. In both situations, soon thereafter one or more parties create and transmit formal "requests for the production of documents" to adversary parties, based on the issues raised in the complaint or letter of inquiry. See the TREC 2006 Legal Track overview for additional background [3].

A survey of case law issued subsequent to the adoption of the new Federal Rules of Civil Procedure in December 2006 suggests that increasing attention is being paid by judges and lawyers to the idea of adversaries in litigation negotiating some form of "search protocol," including coming to consensus on what keywords will be used to search for relevant documents. In one reported case, a judge suggested to the parties that they reach consensus on what form of Boolean queries should be used [13]. In another case, a judge urged the parties to reflect upon recent scholarship discussing the use of "concept searches" to supplement traditional "keyword" searching [7, 9]. Although it remains unclear whether and to what extent lawyers are fully incorporating Boolean and other operators (e.g., proximity operators) in their proposed searches, as an example of best practices the TREC 2007 Legal Track chose to highlight the importance of negotiating Boolean queries by including for each newly created topic a three-stage Boolean query negotiation, consisting of (i) an initial Boolean query[1] as proposed by the receiving party on a discovery request, usually reading the request narrowly; (ii) a "counter"-proposal by the propounding party, usually including a broader set

[1] Although often referred to as "Boolean," these queries contain additional operators (e.g., proximity and truncation operators) that are commonly found in the query languages of commercial search systems that employ Boolean logic.

of terms; and (iii) a final "negotiated" query, representing what was deemed the consensus arrangement as agreed to by the parties without resort to further judicial intervention.

For the TREC 2007 Legal Track, four new hypothetical complaints were created by members of the Sedona Conference® Working Group on Electronic Document Production, a group of lawyers who play a leading role in the development of professional practices for e-discovery. These complaints described: (1) a wrongful death and product liability action based on the use of a certain type of radioactive phosphate resulting in contaminated candy and drinking water; (2) a patent infringement action on a device going by the name "Suck out the Bad, Blow in the Good," designed to ventilate smoke; (3) a shareholder class action suit alleging securities fraud and false advertising in connection with a fictional "Smoke Longer, Feel Younger" campaign relying on "60s-era folk music;" and (4) a fictional Justice Department antitrust investigation looking in to a planned merger and acquisition of a casualty and property insurance company by a tobacco company. As in 2006, in using fictional names and jurisdictions, the track coordinators attempted to ensure that no third party would mistake the academic nature of the TREC Legal Track for an actual lawsuit involving real-world companies or individuals, and any would-be link or association with either past or present real litigation was entirely unintentional.

For each of these four complaints, a set of topics (formally, "requests to produce") were initially created by the creator of the complaint, and revised by the track coordinators. The final topic set contained 50 topics, numbered from 52 to 101. An XML formatted version of the topics (fullL07_v1.xml) was created for (potentially automated) use by the participants.

2.3 Participation

12 research teams participated in this year's Ad Hoc task. The teams experimented with a wide variety of techniques including the following:

- Carnegie Mellon University: structured queries, Indri operators, Dirichlet smoothing, Okapi BM25, boolean constraints, wildcards.
- Dartmouth College: Combination of Expert Opinion (CEO) algorithm, Lemur/Indri, Lucene.
- Fudan University: Indri 2.3, Yatata, word distribution model, corpus pre-processing methods, query expansion, query shrink.
- Open Text Corporation: negotiated boolean queries, defendant boolean, plaintiff boolean, word proximity distances, vector query runs, blind feedback, fusion.
- Sabir Research, Inc.: SMART 16.0, statistical vector space model, ltu.Lnu weighting, Rocchio feedback weighting.
- University of Amsterdam: query formulations, run combinations, LUCENE engine version 1.9, vector-space retrieval model, parsimonious language modeling techniques.
- The University of Iowa (Eichmann): analysis of OCR, 3-4 ngram analysis, translation of boolean query, pseudo-relevance feedback on persons (authors, recipients and mentions) and production boxes.
- The University of Iowa (Srinivasan): Lucene library, Okapi reranking, metadata, wildcard expansion, blind feedback, query reduction.
- University of Massachusetts, Amherst: Indri, term dependence, Markov Random Field (MRF) model, pseudo-relevance feedback, Latent Concept Expansion (LCE), phrase dictionaries, synonym classes, proximity operators.
- University of Missouri, Kansas City: query formulations, vector space model, language model, Lucene, query expansion model, conceptual relevance framework.

- University of Waterloo: Wumpus search engine, cover density ranking, Okapi BM25 ranking, boolean terms, character 4-grams, pseudo-relevance feedback, logistic regression, fusion, CombMNZ combination method, proximity-ranked boolean queries, relaxed boolean.
- Ursinus College: document normalization, log normalization, power normalization, cosine normalization, enhanced OCR error detection, generalized vector space retrieval, query pruning.

The teams submitted a total of 68 experimental runs by the Aug 5, 2007 deadline (each team could submit a maximum of 8 runs). Please consult the individual team papers in the TREC proceedings for the details of the experiments conducted. Also, please check the track web site [1] for the slides of many of the participant presentations at the conference, along with links to the aforementioned individual team members' papers in the TREC proceedings.

2.4 Evaluation

2.4.1 Background on Estimating Precision and Recall

The most straightforward way to produce an unbiased estimate of the number of relevant documents retrieved would be to use simple random sampling (i.e., sampling in which all samples have an equal chance of being selected). Unfortunately, for our purpose, the individual estimates would usually be too inaccurate. For example, suppose the target collection has 7 million documents, and for a particular topic 700 of these are relevant. Suppose further that we have the resources to judge 1,000 documents. If we pick those 1,000 documents from a simple random sample of the collection, most likely 0 of the documents will be judged relevant, producing an estimate of 0 relevant documents, which is far too low. If 1 of the documents were to be judged relevant, then we would produce an estimate of 7,000 relevant documents, which is far too high.

TREC evaluations have typically dealt with this issue by using an extreme variant of stratified sampling. The primary stratum, known as the pool, is typically the set of documents ranked in the top-100 for a topic by the participating systems. Traditionally, all of the documents in the pool are judged. Contrary to the usual approach to stratified sampling, typically none of the unpooled documents are judged (these documents are just assumed non-relevant). For the older TREC collections of about 500,000 documents, [15] found that the results for comparing retrieval systems are reasonably reliable, even though that study also found that probably only 50%-70% of relevant documents for a topic were assessed, on average.

Traditional pooling can be too shallow for larger collections. As the judging pools have become relatively shallower, either from TREC collections becoming larger and/or the judging depth being reduced, concerns have been expressed with the reliability of results. For example, [4] recently reported bias issues with depth-55 judging for the 1 million-document AQUAINT corpus, and [12] estimated that fewer than 20% of the relevant documents were judged on average for the 7 million-document TREC 2006 Legal Track test collection. The TREC 2006 Terabyte Track [5] experimented with taking simple random samples of 200 documents from (up to) depth-1252 pools, and estimated the average precision score for each run based on this deeper pooling by using the "inferred average precision" (infAP) measure suggested by [14]. They found that infAP scores were highly correlated with Mean Average Precision (MAP) scores based on traditional depth-50 pooling.

2.4.2 The L07 Method

The L07 method for estimating recall and precision was based on how the recall and precision components are estimated in the infAP calculation. What distinguishes the L07 method is support for much deeper pooling by sampling higher-ranked documents with higher probability. For legal discovery, recall is of central concern. It was found last year by [12] that marginal precision exceeded 4% on average even at depth 9,000 for standard vector-based retrieval approaches. Hence we used depth-25000 pooling this year to get better coverage of the relevant documents. A simple random sample of a depth-25000 pool, however, would be unlikely to produce accurate estimates for recall at less deep cutoff levels. Hence we sampled higher-ranked documents with higher probability in such a way that recall estimates at all cutoff levels up to max(25000,B) should be of similar accuracy. (Details are provided in the following sections.)

The L07 method was developed independently from the similar "statAP" method evaluated by Northeastern University in the TREC 2007 Million Query Track [2]. (The common ancestor was the infAP method, which also came from Northeastern.) Both methods associate a probability with each document judgment. The differences are in how the probabilities are assigned (which should not matter on average) and in the measures being estimated (we are estimating the recall and precision of a set, whereas statAP is estimating the "average precision" measure which factors in the ranks of the relevant documents). The L07 formulas are provided below, but please consult the Northeastern work for a more thorough discussion of the theoretical underpinnings of measure estimation than we aim to provide here.

2.4.3 Ad Hoc Task Pooling

As stated earlier, a total of 68 runs were submitted by the 12 research teams for the Ad Hoc task by the Aug 5, 2007 deadline. Each run included as many as 25,000 documents (sorted in a putative best-first order) for each of the 50 topics. All submitted runs, plus a 69th run described below, were pooled to depth 25,000 for each topic and then each pool was sampled. The pool sizes before sampling ranged from 195,688 (for topic 76) to 476,252 (for topic 84). (The pool sizes for all of the topics are listed in Section 4.)

The initial plan (given in the Ad Hoc task guidelines) was to assign judging probability p(d) = min(C / hiRank(d), 1) to each submitted document d, where hiRank(d) is the highest (i.e., best) rank at which any submitted run retrieved document d, and C is chosen so that the sum of all p(d) (for all submitted documents d) was the number of documents that could be judged (typically 500). It was hoped that C would be at least 10 for all topics, so that we would have the accuracy of at least 10 simple random sample points for estimates at all depths. After the runs came in, it turned out the C values would range from only 1.6 to 3.3 if judging only 500 documents, substantially limiting the accuracy of the estimates of all measures.

Running some experiments, it turned out for specific depths we could get greater accuracy. For example, if all resources went to a simple random sampling for estimating precision at depth-B, we could get the accuracy of at least 17 sample points for each topic. If instead all resources were directed to just depth-25000, we could get at least 26 sample points for each topic. Of course, if we targeted just one deep measure, we wouldn't have a lot of top-documents for training future systems or for contrasting our measure with traditional rank-based IR measures. Experiments also found that if we just did traditional depth-k pooling, we could only go to at least depth-12 for each topic. But if all resources went to top-12 documents, we wouldn't have the ability to estimate deeper measures.

The sampling process that we ultimately adopted was a hybrid of all of the above. The final p(d) formula for the probability of judging each submitted document d was as follows:

```
If (hiRank(d) <= 5) { p(d) = 1.0; }
Else if (hiRank(d) <= B) { p(d) = min(1.0, ((5/B)+(C/hiRank(d)))); }
Else { p(d) = min(1.0, ((5/25000)+(C/hiRank(d)))); }
```

This formula causes the the first judging bin of 500 documents to contain the top-5 documents from each run, and it causes measures at depths B and 25000 to have the accuracy of approximately 5+C simple random sample points. Measures at other depths will have the accuracy of approximately (at least) C simple random sample points. If C is set to the largest multiple of 0.01 which produces a bin of at most 500 documents, C ranges from 0.34 (topic 82) to 2.42 (topic 76). So by just dropping C by approximately 1 compared to the original plan, we gained more top document judging and at least 5-sample accuracy for depth-B and depth-25000.

To allow for the possibility that some assessors could judge more than 500 documents, the above process was adapted to have a first bin of approximately 500 documents and 5 additional bins of approximately 100 documents each, using the following approach. The C values were set so that the p(d) values would sum to 1,000, and an initial draw of approximately 1000 documents was done. Then the C values were set so that the p(d) values would sum to 900, and approximately 900 documents were drawn from the initial draw of 1000 (using the ratio of the probabilities); the approximately 100 documents that were not drawn became "bin 6". This process was repeated to create "bin 5", "bin 4", "bin 3" and "bin 2". The approximately 500 documents drawn in the last step became "bin 1".

When the judgments were received from the assessors (as described in the next section), the final p(d) values were based on how many bins the assessor had completed (e.g., if 3 bins had been completed, then the p(d) values from choosing C so that the p(d) sum to 700 were used). If there had been partial judging of deeper bins, the judged documents from these bins were also kept, but with their p(d) reset to 1.0. Note that if the 1st bin was not completed, the topic had to be discarded. For each completed topic, the final number of assessed documents and corresponding C values are listed in Section 4.

Two "runs" deserve special mention. First, the reference Boolean run (refL07B), which would have been the 69th run, was not included in the pooling because it had been created by simply resorting one of the pooled runs (otL07fb) alphabetically by docno. Instead, a 69th run called randomL07 was created, which for each topic had 100 randomly chosen documents that were not retrieved by any of the other 68 runs for the topic. We only included 100 random documents per topic, not 25000, to reduce the number of judgments taken away from submitted runs. After the draw, it turned out that the 1st bin of 500 documents to be judged contained between 5 and 15 random documents (average 9.38).

2.5 Relevance Judgments

For the TREC 2007 Legal Track, the track coordinators primarily sought out second-year and third-year law students who would be willing to volunteer as assessors in order to fulfill a law school requirement or expectation to perform some form of pro bono service to the larger community. Based on a nationwide solicitation in mid-August 2007, we received an enthusiastic response from students at a variety of U.S. law schools. All 50 new Ad Hoc task topics for the second year were assigned to assessors, but judgments for 7 topics were not available in time for use in the evaluation.[2] Most of the assessors (42) were law students from a wide variety of institutions: Loyola-L.A. (23 volunteers), University of Indiana-Indianapolis (5), George Washington (3), Case Western Reserve (3), Loyola-New Orleans (2), Boston University (2), University of Dayton (2), University of Maryland (1), and University of Texas (1). Additionally, one Justice Department attorney and one archivist on staff in NARA's Office of the General Counsel participated.

This year, the assessors used a Web-based platform developed by NIST that was hosted at the University of Maryland to view scanned documents and to record their relevance judgments. Assessors found the interface easy to navigate, with the only reported problem being a technical one involving an inability to read or advance screens properly (due to use of a Web browser other than Firefox, the only one supported). Each assessor was given a set of approximately 500 documents to assess, which was labeled "Bin 1." Additional bins 2 through 6, each consisting of 100 documents, were available for optional additional assessment, depending on willingness and time. (It turned out that 8 of the assessors completed at least 1 of the optional bins, and 5 assessors completed all 5 optional bins.) In total, 24,404 judgments were produced for the 43 topics. The assessment phase extended from August 17, 2007 through September 24, 2007.

As in 2006, we provided the assessors with an updated "How To Guide" that explained that the project was modeled on the ways in which lawyers make and respond to real requests for documents, including in electronic form. Assessors were told to assume that they had been requested by a senior partner, or hired by a law firm or another company, to review a set of documents for "relevance." No special, comprehensive knowledge of the matters discussed in each complaint was expected (e.g., no need to be an expert in federal election law, product liability, etc.). The heart of the exercise was to look for relevant and nonrelevant documents within a topic. Relevance, consistent with all known legal definitions from Wigmore to Wikipedia, was to be defined broadly. Special rules were to be applied for any document of over 300 pages. The same process was used for assessment for the interactive and relevance feedback tasks (which had different topics, as described below). See the TREC 2006 Legal Track overview for additional background (including a discussion of inter-assessor agreement which was measured in 2006 but not in 2007) [3].

On the whole, there was less confusion reported by assessors as to the definitional scope of the assigned

[2] The assessments for one additional topic were completed after the deadline, and are available for research use, but results are reported in this paper for 43 topics.

topics in 2007 than in 2006, although some questions did arise. For example, for topic 75 ("All documents that memorialize any statement or suggestion from an elected federal public official that further research is necessary to improve indoor environmental air quality"), the assessor questioned whether "memorialize" would be broad enough to include a mere reference to a Superfund bill, without a quotation as such from the official. We responded that a quotation or allusion to an actual statement made by an official was necessary for the document to be responsive. On the same topic, the assessor wondered if a quote from an appointed federal official (e.g., from the EPA) would qualify, in light of the fact that the negotiated Boolean contained the term "public official" without further qualification. We responded that the topic, not the Boolean string, controlled interpretation, and that the topic contained the additional condition "elected," hence a mere quote from an EPA official without more would not be responsive.

In the case of topic 62, involving press releases concerning water contamination related to irrigation, the assessor reported afterwards that in performing the evaluation "it was sometimes difficult to determine what constituted a press release." Another post-assessment comment stated that because "assessments for responsiveness were done in different sessions, the triggers for responsiveness may not have been consistent," i.e., "sometimes a single word" convinced this assessor that the document was relevant, "while at other intervals I read on to see whether [a finding of relevance] would make more sense in the narrower context of the complaint."

The assessor of topic 80 found it difficult to determine if certain types of radio and magazine advertising were sufficiently clear so as to say that the document made "a connection between folk songs and music and the sale of cigarettes," as the topic required. In the words of the assessor: "While it was easy to identify a connection when a music magazine contained a cigarette ad or when a cigarette magazine contained a music article, other magazines were less obvious. An outdoor magazine[] that contains an interview with a musician as well as a cigarette ad, for example. Or a general interest[] magazine that contains a cigarette ad near its music section." In wondering "how close" the connection had to be, the assessor went on to conclude that "Ultimately, unless the cigarette ad was on the same page as the music section, or in the middle of it, I had to say there was no connection."

One assessor found an error in Complaint C, noting a one-time stray reference to a "Defendant Jones" (at Second Claim for Relief preceding paragraph 46), where all other references in the complaint were to "Defendant Smaug." This circumstance led to a lively debate among track coordinators as to whether the complaint should be left as is, amended for assessors still engaged, or alternatively discarded (we decided to leave it as is given the *de minimis* nature of the error). However, some form of sensitivity analysis might be profitably applied to see if eliminating the anomalous reference changed any run results.

The track coordinators asked assessors to record how much time they spent on their task. Based on 23 survey returns, assessors averaged 25 hours in accomplishing their review of the 500 documents in Bin 1, for an average of 20 documents per assessor per hour. (In 2006 the review rate averaged to 25 documents per hour.) Based on the 2007 returns, the total time devoted works out to approximately 1400 total hours spent on this year's Legal Track tasks (based on 28,141 total judgments divided by 20, including the 24,404 judgments for the 43 Ad Hoc topics, 3,238 judgments for the 10 Interactive/RF topics, and a bin of 499 judgments received after the official results went out). If second-year and third-year law student time were billed at the same rate as summer associates at law firms ($150/hour), those 1400 hours roughly translate to $200,000 in pro bono effort for performing combined relevance assessments during the Ad Hoc and Interactive/RF tasks in 2007.

Not only did the greater cadre this year of law students perform conscientiously during the compressed period of mid-August through mid-September for completing assessments, they appeared to enjoy and benefit from the exercise. Comments from post-assessment surveys included students saying: (i) "On a personal level, the documents were quite interesting. If I had had the time, I gladly would have done another bin of 500, but the semester is starting to get very busy." (ii) "I know more about the effects of cigarettes and smoking than I could have ever thought possible . . ." (iii) "I would love to help out in the future. I found my topic very interesting and enjoyed assessing documents." (iv) "I thought I was getting the short end of the stick because the U.S. Beet Sugar Association had to be the lamest topic of all time. But the documents were really interesting and I learned a lot about the sordid political wrangling over sugar." (v) "I thought that the project was worthwhile from a purely practical standpoint, in that learning how to review massive amounts of information as efficiently as possible is a skill that all lawyers need to work on."

2.6 Results

Each reviewed document was judged relevant, judged non-relevant, or left as "gray." (Our "gray" category includes all documents that were presented to the assessor, but for which a judgment could not be determined. Among the most common reasons for this were documents that were too long to review (more than 300 pages, according to our "How To Guide") or for which there was a technical problem with displaying the scanned document image.)

A qrelsL07.normal file was created in the common trec_eval qrels format. Its 4th column was a 1 (judged relevant), 0 (judged non-relevant), -1 (gray) or -2 (gray). (In the assessor system, -1 was "unsure" (the default setting for all documents) and -2 was "unjudged" (the intended label for gray documents).)

A qrelsL07.probs file was also created, which was the same as qrelsL07.normal except that there was a 5th column which listed the p(d) for the document (i.e., the probability of that document being selected for assessment from the pool of all submitted documents). qrelsL07.probs can be used with the experimental l07_eval utility to estimate precision and recall to depth 25,000 for runs which contributed to the pool.

2.6.1 Estimating the Number of Relevant Documents for a Topic

To estimate the number of relevant, non-relevant and gray documents in the pool for each topic, the following procedure was used:

Let D be the set of documents in the target collection. For the Legal Track, $|D|$=6,910,912.

Let S be a subset of D.

Define $JudgedRel(S)$ to be the set of documents in S which were judged relevant.

Define $JudgedNonrel(S)$ to be the set of documents in S which were judged non-relevant.

Define $Gray(S)$ to be the set of documents in S which were reviewed but not judged relevant nor non-relevant.

Define $estRel(S)$ to be the estimated number of relevant documents in S:

$$estRel(S) = \min\left(\left(\sum_{d \in JudgedRel(S)} \frac{1}{p(d)}\right),\ (|S| - |JudgedNonrel(S)|)\right) \tag{1}$$

Note that $estRel(S)$ is 0 if $|JudgedRel(S)| = 0$.

Define $estNonrel(S)$ to be the estimated number of non-relevant documents in S:

$$estNonrel(S) = \min\left(\left(\sum_{d \in JudgedNonrel(S)} \frac{1}{p(d)}\right),\ (|S| - |JudgedRel(S)|)\right) \tag{2}$$

Note that $estNonrel(S)$ is 0 if $|JudgedNonrel(S)| = 0$.

Define $estGray(S)$ to be the estimated number of gray documents in S:

$$estGray(S) = \min\left(\left(\sum_{d \in Gray(S)} \frac{1}{p(d)}\right),\ (|S| - (|JudgedRel(S)| + |JudgedNonrel(S)|))\right) \tag{3}$$

Note that $estGray(S)$ is 0 if $|Gray(S)| = 0$.

Applying the above formulas, the estimated number of relevant documents in the pool, on average per topic, was 16,904(!). The number varied considerably by topic, from 18 (for topic 63) to 77,467 (for topic 71). (The estimates for all of the topics are listed in Section 4.) Obviously, traditional top-ranked pooling would not have been sufficient to cover the high numbers of relevant documents. On average (per topic), the estimated number of non-relevant documents in the pool was 298,678, and the estimated number of gray documents in the pool was 4,303.

2.6.2 Estimating Recall

The L07 approach to estimating recall is similar to how infAP estimates its recall component (i.e., it's based on the run's judged relevant documents found to depth k compared to the total judged relevant documents). In our case, we have to weight the judged relevant documents to account for the sampling probability.

For a particular ranked retrieved set S:

Let $S(k)$ be the set of top-k ranked documents of S. (Note that $|S(k)| = \min(k, |S|)$.)

Define *estRecall@k* to be the estimated recall of S at depth k:

$$estRecall@k = \frac{estRel(S(k))}{estRel(D)} \tag{4}$$

The mean estimated recall of the reference Boolean run (refL07B) was just 22%. Hence the final negotiated boolean query was missing about 78% of the relevant documents (on average across all topics). Note that the estimated recall of refL07B varied considerably per topic, from 0% (topic 77) to 100% (topic 84). Figure 1 illustrates the breakdown of the estimated relevant documents into those matching the Boolean query and those only found by at least one of the ranked systems.

Section 4 lists the final Boolean query's estimated recall for each topic; it also lists several relevant documents (underlined) which did not match the final Boolean query. For example, it shows that for topic 74 ("All scientific studies expressly referencing health effects tied to indoor air quality") the final negotiated Boolean query of "`(scien! OR stud! OR research) AND ("air quality" w/15 health)`" missed matching relevant document vkm92d00. This document did not contain required Boolean terms such as "air" or "health", but it was judged relevant presumably because it referred to the "largest study ever" on whether "secondary smoke causes cancer" and also to a study of the "carcinogenic effects" of the "gas released from volatile organic compounds in shower water"[3].

Despite the low recall of the final Boolean query, none of the 68 submitted runs had a higher mean estimated Recall@B than the reference Boolean run (as Table 1 shows). This is a surprising result, since the refL07B run had been available to the participants since early July and thus could have been used by participating systems as one source of evidence. At least one participant (Open Text, which contributed the reference Boolean run) reported that combining other techniques with the Boolean run did not increase average recall at depth B. We anticipate that more participants will attempt to make use of the reference Boolean run next year.

One factor that may be limiting average recall across topics is that our Boolean queries targeted a B range between 100 and 25,000 to keep the size of submitted runs manageable. We should perhaps review whether our Boolean queries might be higher precision (and hence lower recall) than the Boolean queries used in practice.

Table 1 also lists mean estimated Recall@25000. The highest score (47%) was by a run which used both the final Boolean and request text fields (wat1fuse).

Section 4 lists the median scores for each topic in both Recall@B and Recall@25000. Although the final Boolean query had a higher recall than the median Recall@B for 31 of the 43 topics (and 4 tied), the median Recall@25000 was higher than the Boolean query's recall for 33 of the 43 topics (and 1 tied). The median run typically could not match the precision of the Boolean query to depth B, but by retrieving deeper it typically would find more relevant documents. (Table 1 shows that the average B value was 5004, while most of the runs retrieved the allowed maximum of 25,000 per topic.)

2.6.3 Estimating Precision

The L07 approach to estimating precision is similar to how infAP estimates its precision component (i.e., it's based on the precision of the run's judged documents to depth k). In our case, we have to weight the judged documents to account for the sampling probability. We also multiply by a factor to ensure that unretrieved documents are not inferred to be relevant.

[3] In the OCR output used by the participants, this latter phrase actually appeared as "Ras released f rom volatile organtc compounds in .houer water".

		Avg.	Est.	Est.	Est.	Est.	Est.	Est.			(raw)
Run	Fields	Ret.	R@B	R25000	P5	P@B	P25000	Gray@B	S1J	GS10J	R-Prec
refL07B	bM	5004	**0.216**			**0.292**		0.042			
otL07fb	bM	5004	**0.216**	0.216	0.507	**0.292**	0.056	0.042	24/43	0.863	0.201
otL07fb2x	bM	6053	0.209	0.242	0.486	0.282	0.065	0.031	21/43	0.837	0.193
CMUL07ibs	bM	25000	0.208	0.392	0.452	0.267	0.137	0.022	24/43	0.832	0.151
wat5nofeed	brM	24999	0.196	0.447	0.537	0.263	0.159	0.016	24/43	0.864	0.204
CMUL07irs	bM	25000	0.194	0.395	0.372	0.242	0.149	0.013	23/43	0.813	0.133
otL07frw	brM	25000	0.193	0.428	**0.550**	0.278	0.150	0.015	**26/43**	**0.883**	**0.224**
CMUL07ibt	bM	25000	0.187	0.391	0.495	0.261	0.136	0.024	23/43	0.847	0.175
otL07pb	pM	18555	0.186	0.327	0.424	0.235	0.119	0.025	17/43	0.792	0.147
wat1fuse	brM	25000	0.186	**0.469**	0.529	0.271	0.155	0.012	21/43	0.837	0.197
CMUL07ibp	bM	25000	0.183	0.392	0.472	0.252	0.131	0.026	**26/43**	0.859	0.178
wat7bool	bM	7059	0.172	0.198	0.420	0.250	0.064	0.047	18/43	0.761	0.140
CMUL07o3	bdpM	25000	0.170	0.400	0.491	0.236	0.146	0.009	23/43	0.867	0.190
otL07fbe	bM	25000	0.169	0.369	0.492	0.273	0.160	0.015	22/43	0.792	0.178
IowaSL0704	bdpr	20071	0.167	0.382	0.461	0.232	0.123	0.015	21/43	0.760	0.173
UMass15	r	25000	0.165	0.354	0.465	0.229	0.122	0.006	23/43	0.789	0.172
IowaSL0703	bdpr	25000	0.164	0.403	0.451	0.216	0.139	0.009	19/43	0.808	0.181
otL07fv	bM	25000	0.163	0.357	0.467	0.225	0.129	0.005	20/43	0.856	0.176
UMass13	br	25000	0.162	0.322	0.442	0.195	0.118	0.007	20/43	0.861	0.175
IowaSL0705	dpr	7396	0.161	0.260	0.502	0.243	0.066	0.017	22/43	0.799	0.184
UMKC4	d	24419	0.157	0.412	0.391	0.253	0.145	0.014	16/43	0.749	0.144
IowaSL0706	br	7396	0.153	0.247	0.428	0.252	0.067	0.012	21/43	0.792	0.176
otL07rvl	rM	25000	0.153	0.420	0.471	0.235	0.151	0.010	15/43	0.850	0.187
CMUL07o1	bM	25000	0.152	0.361	0.467	0.204	0.133	0.009	24/43	0.848	0.168
UMass12	b	24028	0.150	0.285	0.350	0.197	0.117	0.005	21/43	0.804	0.125
SabL07arbn	bdpr	25000	0.149	0.321	0.393	0.203	0.117	0.009	19/43	0.790	0.132
UMass14	r	25000	0.147	0.362	0.464	0.208	0.113	0.010	25/43	0.813	0.167
wat8gram	bM	25000	0.142	0.389	0.419	0.256	0.143	0.009	21/43	0.781	0.137
IowaSL0707	brM	5004	0.142	0.142	0.409	0.237	0.053	0.013	20/43	0.781	0.173
wat6qap	bM	17179	0.142	0.239	0.385	0.194	0.089	0.031	13/43	0.733	0.113
UMKC6	b	25000	0.137	0.400	0.371	0.258	0.161	0.020	20/43	0.794	0.149
UMass11	r	25000	0.137	0.325	0.377	0.191	0.104	0.007	14/43	0.764	0.145
UMKC1	d	24419	0.135	0.426	0.436	0.241	0.153	0.019	20/43	0.752	0.124
IowaSL0702	bdpr	25000	0.134	0.363	0.429	0.205	0.132	0.007	19/43	0.816	0.183
SabL07ab1	bdpr	25000	0.132	0.316	0.371	0.204	0.123	0.013	18/43	0.747	0.119
CMUL07irt	bM	25000	0.132	0.294	0.467	0.189	0.115	0.013	22/43	0.829	0.158
UMass10	rM	23649	0.131	0.306	0.489	0.207	0.117	0.016	23/43	0.800	0.136
UMKC5	r	25000	0.126	0.411	0.370	0.260	**0.172**	0.012	18/43	0.750	0.148
wat3desc	rM	24999	0.124	0.394	0.426	0.235	0.143	0.010	20/43	0.775	0.160
CMUL07std	rM	25000	0.123	0.314	0.451	0.191	0.115	0.006	22/43	0.833	0.163
fdwim7xj	rM	25000	0.113	0.354	0.408	0.180	0.114	0.017	22/43	0.781	0.142
ursinus1	r	25000	0.113	0.329	0.340	0.195	0.125	0.016	16/43	0.751	0.131
ursinus2	r	25000	0.112	0.314	0.307	0.154	0.117	0.008	14/43	0.685	0.099
ursinus6	r	25000	0.110	0.298	0.242	0.153	0.108	0.009	10/43	0.628	0.089
IowaSL07Ref	r	25000	0.108	0.343	0.366	0.148	0.120	0.009	15/43	0.754	0.130
UMKC3	b	25000	0.107	0.391	0.444	0.243	0.161	0.031	17/43	0.790	0.131
fdwim7rs	r	25000	0.106	0.319	0.431	0.210	0.126	0.013	21/43	0.790	0.142
UIowa07LegE2	b	16708	0.106	0.268	0.283	0.224	0.114	0.021	11/43	0.651	0.109
UIowa07LegE0	r	24997	0.103	0.318	0.312	0.156	0.120	0.009	19/43	0.736	0.118
fdwim7ss	cir	25000	0.102	0.309	0.409	0.170	0.115	0.014	17/43	0.788	0.129
fdwim7sl	cir	25000	0.101	0.288	0.422	0.164	0.120	0.010	20/43	0.783	0.120
UMKC2	r	25000	0.100	0.409	0.416	0.226	0.171	0.016	17/43	0.770	0.125
ursinus7	bM	25000	0.099	0.283	0.265	0.161	0.109	0.010	12/43	0.601	0.086
SabL07ar2	r	25000	0.098	0.295	0.364	0.178	0.105	0.016	16/43	0.789	0.102
SabL07ar1	r	25000	0.097	0.288	0.369	0.174	0.105	0.015	15/43	0.784	0.101
ursinus4	r	25000	0.096	0.315	0.332	0.233	0.139	**0.131**	14/43	0.714	0.066
Dartmouth1	r	25000	0.083	0.275	0.285	0.137	0.102	0.010	15/43	0.682	0.095
wat2nobool	brM	25000	0.082	0.320	0.327	0.217	0.131	0.018	12/43	0.713	0.071
ursinus8	bM	25000	0.071	0.191	0.093	0.101	0.085	0.018	4/43	0.416	0.032
fdwim7ts	r	25000	0.070	0.177	0.163	0.109	0.067	0.012	6/43	0.592	0.044
ursinus3	r	25000	0.063	0.213	0.072	0.084	0.078	0.008	3/43	0.401	0.024
ursinus5	r	25000	0.063	0.220	0.072	0.081	0.083	0.009	3/43	0.396	0.023
wat4feed	brM	25000	0.061	0.224	0.261	0.151	0.092	0.025	9/43	0.600	0.063
catchup0701p	r	24016	0.061	0.171	0.126	0.130	0.098	0.023	5/43	0.528	0.041
UIowa07LegE1	r	24996	0.031	0.083	0.206	0.067	0.057	0.004	11/43	0.651	0.036
UIowa07LegE3	b	24999	0.028	0.110	0.194	0.090	0.070	0.012	9/43	0.550	0.035
otL07db	dM	368	0.027	0.027	0.301	0.026	0.006	0.003	15/43	0.576	0.074
UIowa07LegE5	b	24992	0.003	0.019	0.105	0.018	0.026	0.017	6/43	0.324	0.011
UIowa07LegE4	b	9879	0.002	0.004	0.023	0.012	0.009	0.010	1/43	0.101	0.002
randomL07		100	0.000	0.000	0.000	0.000	0.000	0.001	0/43	0.038	0.001

Table 1: Mean scores for submitted Ad Hoc task runs.

Run	Depths 1-5000	Depths 5001-10000	Depths 10001-15000	Depths 15001-20000	Depths 20001-25000
Ad Hoc - high	0.238	0.182	0.183	0.168	0.199
Ad Hoc - median	0.178	0.127	0.110	0.101	0.099
RF - high	0.218	0.197	0.146	0.197	0.150
RF - median	0.126	0.118	0.069	0.055	0.052

Table 2: High and Median Estimated Marginal Precision Rates

Define $estPrec@k$ to be the estimated precision of S at depth k:

$$estPrec@k = \frac{estRel(S(k))}{estRel(S(k)) + estNonrel(S(k))} \times \frac{|S(k)|}{k} \quad (5)$$

Note: we define $estPrec@k$ as 0 if both $estRel(S(k))$ and $estNonrel(S(k))$ are 0.

The mean estimated precision of the reference Boolean run (refL07B) was just 29%. Again, this varied by topic, from 0% (topic 77) to 97% (topic 69) (Section 4 lists the precision scores for all of the topics). As Table 1 shows, none of the 68 submitted runs had a higher mean estimated Precision@B than the reference Boolean run. The submitted run with the highest mean estimated Precision@25000 (17%) just used the request text field (UMKC5).

2.6.4 Marginal Precision Rates to Depth 25,000

Table 2 shows how precision falls with retrieval depth for the Ad Hoc task runs. The table includes the highest and median estimated marginal precision rates of the Ad Hoc task runs for depths 1-5000, 5001-10000, 10001-15000, 15001-20000 and 20001-25000. The median run was still maintaining 10% precision at the deepest stratum (depths 20001-25000), and some runs were close to 20% precision in this stratum. It appears for this test collection that depth 25,000 was not deep enough to cover all of the relevant documents that a run could potentially find.

We note however that the median precision in the deepest stratum exceeded 10% for only 6 of the 43 topics. These included topics 69, 74 and 71, which also had among the highest number of estimated relevant documents (as per the listing in Section 4). 13 of the 43 topics had more than 25,000 estimated relevant documents (hence 100% recall was not possible for these topics at depth 25,000). Perhaps it would be better for reusability to discard these 13 topics, but additional analysis will be needed before we can draw firm conclusions.

We hope to investigate reusability issues with the collection as (near) future work. But generally speaking, for runs that did not contribute to the pools, if the 25,000 documents retrieved for a topic are mostly a subset of the (approximately) 300,000 documents that were pooled for the topic, or if the unpooled documents contain few relevant documents, then the estimated measures from the l07_eval utility should still be comparable to the pooled systems' scores, particularly at deeper depths (e.g., at depth 25,000).

2.6.5 Estimated Gray Percentages

Table 1 also lists the estimated percentage of gray documents at depth B. The ursinus4 run retrieved a lot more gray documents (13%) than the other runs; we therefore suspect this run's approach favored longer documents. Boolean runs retrieved 4-5% gray, perhaps because the Boolean constraints matched some long documents. Most other runs retrieved less than 2% gray. These systematic differences suggest that it may be productive to reconsider the techniques being used for dealing with long documents, both in system design and in our assessment process.

2.6.6 Random Run Results

It was hoped that the randomL07 run would give us at least a rough indication of the number of relevant documents that may have been outside of the pooled results of participating systems. A total of 446 of the randomL07 documents were judged (over 43 topics), and 3 of these were judged relevant (0.7%). Typically, about 6.5 million documents of the almost 7 million documents in the collection were not submitted by any participating system and thus not included in our pooling process. If 0.7% of those are relevant, that would suggest another 50,000 relevant documents (per topic). However, when we reviewed the 3 judged relevant documents from the randomL07 run (zsm67e00 for topic 58, cdf53e00 for topic 70 and dkb74d00 for topic 71), none of them appeared to actually be relevant to us (though the official qrels have not been altered). So perhaps the overall precision of the unpooled documents is actually much less than 0.7% (though even 0.1% would represent more than 5,000 relevant documents per topic).

There were 6 randomL07 documents that were considered "gray" (i.e., not judged relevant nor non-relevant). None of these 6 documents were very long. For one of them, the PDF document would not display (bev71d00 for topic 84). 2 of the 6 were non-English documents (tpv77e00 and xyu37e00 for topic 52). The other 3 had relatively little text (lkf03d00 and xqx21a00 for topic 69, and qge12c00 for topic 89). However, these documents do not seem to be typical of the gray documents retrieved by system runs, which generally were very long documents.

2.6.7 Table Glossary

Table 1 lists the mean scores for each of the 68 submitted runs for the Ad Hoc task and the 2 additional reference runs. The following glossary explains the codes used in that table.

"Fields": The topic fields used by the run: 'b' Boolean query (final negotiated), 'c' complaint, 'd' defendant Boolean (initial proposal), 'i' instructions and definitions, 'p' plaintiff Boolean (rejoinder query), 'r' request text, 'M' manual processing was involved, 'F' feedback run (2006 relevance assessments were used, applicable to RF task only).

"Avg. Ret.": The Average Number of Documents Retrieved per Topic.

"R@B" and "R@25000": Estimated Recall at Depths B and 25000.

"P5", "P@B" and "P25000": Estimated Precision at Depths 5, B and 25000.

"Gray@B": Estimated percentage of Gray documents at Depth B.

"S1J": Success of the First Judged Document.

"GS10J": Generalized Success@10 on Judged Documents (1.08^{1-r} where r is the rank of the first relevant document, only counting judged documents, or zero if no relevant document is retrieved). GS10J is a robustness measure which exposes the downside of blind feedback techniques [11]. Intuitively, it is a predictor of the percentage of topics for which a relevant document is retrieved in the first 10 rows.

"R-Prec": R-Precision (raw Precision at Depth R, where R is the raw number of known relevant documents). Estimation is not used for this measure. It is provided so that we can see if the results differ with a traditional IR measure.

For the reference Boolean run (refL07B), only measures at depth B are shown so that the ordering of Boolean results does not matter.

3 Interactive and Relevance Feedback Tasks

In 2006, most teams applied existing information retrieval systems to obtain what might best be characterized as baseline results. Moreover, in 2006 the relevance assessment pools were (with two exceptions) drawn from near the top of submitted ranked retrieval runs. Both factors tend to reduce the utility of the available relevance judgments for the 2006 topic set somewhat. We therefore created two opportunities for teams to contribute runs that would permit us to enrich the 2006 relevance judgments: a Relevance Feedback task, and an Interactive task. The same document collection was used in 2006 and 2007, so participation in these tasks did not require indexing a new collection.

The objective in the Relevance Feedback task was to automatically discover previously unknown relevant documents by augmenting the evidence available from the topic description with evidence available from the relevance assessments that were created in 2006. Teams could use positive and/or negative judgments in conjunction with the metadata for and/or full text from the judged documents to refine their models. This task provides a simple and well controlled model for assessing the utility of a two-pass search process.

Interactive searchers have an even broader range of strategies available for enhancing their results, including iteratively improving their query formulation (based on their examination of search results) and/or performing more than one iteration of relevance feedback. There is, therefore, significant scope for research on processes by which specific search technologies can best be employed. Participants in the Interactive task could use any combination of: systems of their own design, the Legacy Tobacco Document Library system (LTDL, a Web-based system provided by the University of California, San Francisco), or the Tobacco Documents Online system (TDO, the same Web-based system that was used for relevance assessment in the TREC 2006 Legal Track). Standardized questionnaires were used to collect information about the search process from the perspective of individual participants, and research teams could employ additional methods (e.g., observation or log file analysis) to collect complementary information at their option.

3.1 Topics

Twelve topics out of the first year's 39 completed topics were selected for the Interactive task. These topics were chosen by the track coordinators based on a variety of factors, including (i) not being too closely tied to a "tobacco-related" topic, so as to mitigate whatever inherent bias exists; (ii) the absolute number of relevant documents found for each topic in year one, with topics returning under a threshold of 50 documents being considered lesser priority; (iii) relatively high kappa scores from year 1 on inter- assessor agreement, and (iv) their inherent interest. The top three interactive topics ended up, in priority order, involving the subjects of pigeon deaths (topic 45), memory loss (topic 51), and the placement of tobacco products in G-rated movies (topic 7). A total of 10 of the 12 interactive topics were completely assessed by volunteers. These same 10 topics were used for the Relevance Feedback task in 2007.

3.2 Evaluation

Eight Relevance Feedback runs were submitted by 3 research teams. Participating teams were allowed to submit up to max(25000,B)+1000 documents per topic. "Residual evaluation" was used for the Relevance Feedback task. Hence, before pooling, any documents that were judged last year (of which there were at most 1000 per topic) were removed from the Relevance Feedback runs. For topics with B_r>25000, which was the case for two of the Relevance Feedback topics, depth-B_r pooling was used; other topics were pooled to depth 25,000. The resulting ranked lists were therefore truncated at max(25000,B_r), where B_r is the number of documents matching the reference Boolean query (refL06B) after last year's judged documents are removed. Because B_r>25000 for two topics, R@B_r scores can exceed R@25000 in the Relevance Feedback task, which is not possible for the Ad Hoc task.

The pools were then enriched before judgment with three additional runs:

- A special "oldrel07" run was added which included 25 relevant documents (or all if less than 25 were available) randomly chosen from last year's judgments for each topic.
- A special "oldnon07" run was added which included 25 non-relevant documents randomly chosen from last year's judgments for each topic.
- A special "randomRF07" run was added which included 100 randomly chosen documents that were not otherwise pooled (or judged last year).

Finally, all documents from the Interactive task runs were included in this year's pools (even if they were judged in 2006).

The p(d) formula for the Relevance Feedback task was the same as for the Ad Hoc task except that: (1) all documents from the Interactive task and from the oldrel07 and oldnon07 runs were assigned probability

1.0 so that they would all be presented to the assessor, (2) topics with B_r>25000 were sampled to depth B_r, with p(d) of min(1.0, ((5/25000)+(C/hiRank(d))) for documents with hiRank(d) between 6 and 25000 inclusive, and (3) the sum of the p(d) could be just 250 instead of 500 for some of the topics (because fewer documents needed to be judged to maintain the same accuracy (C value) as in the Ad Hoc task).

For the Interactive task, teams could submit as many as 100 documents per topic, but submission of nonrelevant documents was penalized. A simple utility function (the number of submitted relevant documents minus half the number of submitted nonrelevant documents) was chosen as the principal evaluation measure for the Interactive task in order to encourage participants to submit fewer than 100 documents when that many relevant documents could not be found.

3.3 Interactive Task Results

Table 3 shows the results for each Interactive task team. Eight teams from three sites participated:

Long Island University (LIU). Nine participants worked in groups of three, with one group assigned to each of the highest priority topics. All three groups searched only the LTDL. A tenth participant reviewed each group's top 100 retrieved results and only those considered relevant were submitted. The estimated total search time (across all three searchers in a team) was 39 hours for topic 51, 39 hours for topic 45, and 25 hours for topic 7. Results were reported as both Bates numbers and URL's; the DOCNO was extracted from the URL for pooling and scoring. The reported results for LIU were corrected after they were initially distributed to remove 14 LTDL documents that are not contained in the TREC Legal Track test collection.

Sabir Research (Sabir). One participant worked alone to search the eight highest priority topics. Multiple relevance feedback iterations were performed based on judgments from 2006, plus judgments for an additional 10 previously unjudged documents that were added at each iteration. The limited multi-pass interaction in this process was intended as a contrastive condition to the one-pass relevance feedback runs from the same site; comparison with results from other sites may be less informative because manual query refinement was not performed in this case. The process required an average of about 16 minutes per topic. Results were reported as both Bates numbers and URL; the DOCNO was therefore extracted from the URL.

University of Washington (UW). Sixteen participants worked in six teams, each consisting of two or three participants, and the results for each team were submitted and scored separately. The average time per topic was not reported. Results were submitted as Bates numbers and were automatically mapped to DOCNO values based on exact string match. This process resulted in frequent failures (44% of all values reported as Bates numbers proved to be unmappable); inspection of the values revealed that some were very similar to valid Bates numbers that were present in the collection (e.g., with a dash in place of a slash or with a brief prefix indicating the source), but that others were not. After relevance judgments were completed, the mapping script was modified to accommodate some patterns that were detected by inspection and the UW runs were rescored. The rescored values are shown in italics in Table 3 where they differ from the initially computed (completely assessed) results.

The evaluation design that we chose can in some sense be thought of as comparing the opinion of one person (the relevance assessor) with the opinion of one or more other people (the experiment participants). From the LIU results, we can see that a moderately good level of agreement can be achieved, with agreement on 96 (81%) of the 118 positive judgments made by the participants. Perhaps the most interesting conclusion that we can draw from comparing the results from the seven LIU and UW teams that tried a fully interactive process is that searchers exhibited substantial variation. For example, among the six UW teams (which were given consistent instructions) we see a variation of at least a factor of two in the number of relevant documents found (in the opinion of the relevance assessor) for each of the three topics. Of course, we must caveat this result by noting that the process used for recording Bates numbers by UW participants exhibited substantial variation both by team and by topic, so variations in the effectiveness of the mapping process are a possible confounding factor.

Team	Score	45S	45R	45N	51S	51R	51N	7S	7R	7N
LIU	**85.0**	13.5	18	9	20.0	24	8	**51.5**	**54**	5
UW6	*66.5*	**28.0**	**35**	14	*15.0*	*19*	*8*	23.5	38	29
UW2	*51.5*	*24.5*	32	*15*	15.5	31	31	*11.5*	*24*	25
UW1	48.0	17.0	28	22	**24.0**	**35**	22	7.0	10	6
UW3	*43.0*	*16.0*	*17*	*2*	9.5	23	27	17.5	30	25
UW5	*39.0*	*12.5*	21	*17*	*15.0*	23	*16*	11.5	23	23
UW4	*36.0*	*18.0*	*28*	*20*	*5.5*	*17*	*23*	*12.5*	20	*15*
Sabir	-10.5	-7.0	11	36	-0.5	0	1	3.0	3	12
oldrel07	14.5	5.5	12	13	12.0	16	8	-3.0	6	18
refL06B	12.0	6.5	71	129	7.0	73	132	-1.5	16	35
randomRF07	-23.5	-10.0	0	20	-5.0	1	12	-8.5	0	17
oldnon07	-34.5	-11.5	0	23	-10.5	1	23	-12.5	0	25

Table 3: Mean scores for all Interactive task teams (S=Score, R=relevant, N=Not relevant).

Table 3 also lists the scores of the reference Boolean run (refL06B). Most of the Interactive teams scored higher than the Boolean run on all 3 topics. It should be noted that, unlike the participants, the Boolean run was not limited to 100 retrieved documents, and its results were sampled unevenly (it includes the documents selected for the residual RF evaluation along with the documents from the Interactive, oldrel07 and oldnon07 runs).

3.4 Relevance Feedback Task Results

Table 4 shows the results for the 10 Relevance Feedback runs. Three research teams participated, and two additional reference reference runs were also scored (refL06B and randomRF07). The following summarizes each team's submissions:

Carnegie Mellon University (CMU07). The CMU team treated the Relevance Feedback task as a supervised learning problem and retrieved documents using Indri queries that approximated Support Vector Machines (SVMs) learned from training documents. Both keyword features and simple structured "term.field" features were investigated. Named-Entity tags, the LingPipe sentence breaker, and metadata provided in the collection with each document were used to generate the field information. The CMU07RSVMNP run included negative and positive weight terms, while the CMURFBSVME run was formulated using only terms with positive weights.

Open Text Corporation (ot). The two runs submitted by Open Text performed the Relevance Feedback task, but without actually using the available positive or negative assessments. Run otRF07fb ranked the documents in the reference Boolean run (refL06B). Run otRF07fv performed ranked retrieval using the query terms from the same Boolean query.

Sabir Research (sab07). Sabir's runs provided a baseline for multi-pass relevance feedback runs that were submitted for the Interactive task.

Mean scores over just 10 topics may not be very reliable, so little should be read from the result that no participating system outperformed the reference Boolean run on the mean Est. R@B_r measure or that every run (other than the random run) outperformed the reference Boolean run on the mean Est. P@B_r measure. An encouraging result from this pilot study is that the median of the 5 feedback runs outscored the reference Boolean run in Est. R@B_r for 7 of the 10 topics, in contrast with the Ad Hoc task in which the median run outscored the reference Boolean run in just 8 of the 43 topics. Of course, these results were obtained on different topics, by different numbers of teams, and 10 topics remains a small sample however you slice it (the track hopes to conduct a larger study in the upcoming year). Some useful insights may come from failure analysis of the topics for which the feedback runs were still outscored by the reference Boolean run.

		Avg.	Est.	Est.	Est.	Est.	Est.	Est.			(raw)
Run	Fields	Ret.	R@B$_r$	R25000	P5	P@B$_r$	P25000	Gray@B$_r$	S1J	GS10J	R-Prec
refL06B	bM	20185	**0.386**			0.106		**0.051**			
otRF07fb	bM	20185	**0.386**	0.367	0.380	0.106	0.044	**0.051**	3/10	0.735	0.100
CMU07RSVMNP	F-bM	25159	0.380	0.598	**0.570**	0.170	**0.155**	0.007	5/10	**0.870**	0.182
CMU07RBase	bM	25100	0.353	0.578	0.500	0.115	0.102	0.024	**7/10**	0.843	0.117
CMU07RFBSVME	F-bM	25116	0.334	0.578	**0.570**	0.118	0.072	0.034	6/10	0.855	0.173
sab07legrf2	F-bdpr	36625	0.333	0.515	0.500	0.173	0.099	0.012	4/10	0.788	0.199
sab07legrf3	F-bdpr	36625	0.321	**0.616**	0.520	**0.194**	0.117	0.012	4/10	0.814	**0.202**
sab07legrf1	F-brM	25106	0.278	0.475	0.480	0.132	0.072	0.013	5/10	0.802	0.197
otRF07fv	bM	36625	0.248	0.444	0.355	0.156	0.064	0.007	3/10	0.612	0.065
randomRF07		100	0.002	0.002	0.040	0.004	0.000	0.000	0/10	0.159	0.004

Table 4: Mean scores for submitted Relevance Feedback task runs.

4 Individual Topic Results

This section provides summary information for each of the assessed topics of the Ad Hoc and Relevance Feedback tasks.

The 44 assessed Ad Hoc topics are listed first (including one topic (#62) whose judgments arrived after the official results had been released). The information provided is as follows:

Topic : The topic numbers range from 52 to 101. In parentheses are the year of the topic set (2007), the label of the complaint (A, B, C or D), and the request number inside the complaint. (The complaints run several pages and are available on the track web site [1].)

Request Text : The one-sentence statement of the request.

Initial Proposal by Defendant, Rejoinder by Plaintiff and Final Negotiated Boolean Query :

The following syntax was used for the Boolean queries:

- `AND, OR, NOT, ()`: As usual.
- `x BUT NOT y`: Same meaning as `(x AND (NOT (y)))`
- `x`: Match this word exactly (case-insensitive).
- `x!`: Truncation – matches all strings that begin with substring `x`.
- `!x`: Truncation – matches all strings that end with substring `x`.
- `x?y`: Single-character wildcard – matches all strings that begin with substring `x`, end with substring `y`, and have exactly one-character in between `x` and `y`.
- `x*y`: Muliple-character wildcard – matches all strings that begin with substring `x`, end with substring `y`, and have 0 or more characters between `x` and `y`.
- `"x"`, `"x y"`, `"x y z"`, etc.: Phrase – match this string or sequence of words exactly (case-insensitive).
- `"y x!"`, `"x! y"`, etc.: If `!` is used internal to a phrase, then do the truncated match on the words with `!`, and exact match on the others. (The `*` and `?` wildcard operators may also be used inside a phrase.)
- `w/k`: Proximity – `x w/k y` means match `"x a b ... c y"` or `"y a b ... c x"` if `"a b ... c"` contains `k` or fewer words.
- `x w/k1 y w/k2 z`: Chained proximity – a match requires the same occurrence of `y` to satisfy `x w/k1 y` and `y w/k2 z`.

Sampling and Est. Rel. : The number of pooled documents is given (i.e., all distinct documents from all of the submitted runs for the topic), followed by the number presented to the assessor, the number the assessor judged relevant, the number the assessor judged non-relevant, and the number the assessor left as "gray" (defined earlier). The "C" value of the p(d) formula is given (the measures at depths B and 25000 have approximately the accuracy of 5+C simple random sample points, as described earlier). "Est. Rel." is the estimated number of relevant documents in the pool for the topic (based on the sampling results).

Final Boolean Result Size (B), Est. Recall and Est. Precision : "B" is the number of documents matching the final negotiated Boolean query (which always was between 100 and 25000 in 2007). "Est. Recall" is the estimated recall of the final negotiated Boolean query result set. "Est. Precision" is the estimated precision of the final negotiated Boolean query result set.

Participant High Recall@B and Median Recall@B : The highest estimated recall at depth B of the participant runs is listed, followed in parentheses by the run's identifier; if more than one run tied for the highest score, just one of them is listed (chosen randomly) and the number of other tied runs is stated. The median estimated recall at depth B is based on the median score of 70 runs (the 68 participant runs along with the refL07B and randomL07 runs).

Participant High Recall@25000 and Median Recall@25000 : Same as previous line except that the measures are at depth 25000 instead of depth B.

5 Deepest Sampled Relevant Documents : The identifiers of the 5 deepest sampled documents that were judged relevant are listed in descending depth order. The identifier is underlined if the document was not retrieved by the final negotiated Boolean query. Each identifier is followed by its weight in the estimation formulas (i.e., the estimated number of relevant documents it represents) which is 1/p(d) (i.e., the reciprocal of the probability of being selected for judging). In parentheses is the identifier of the run which retrieved the document at the highest rank, followed by that rank (which can range from 1 to 25000); if multiple runs retrieved the document at that rank, just one of them is listed. For example, for topic 52, the entry of "hdz83f00-93.4 (otL07fbe-515)" indicates that the hdz83f00 document was judged relevant and, because it is not underlined, it matched the final Boolean query. It counts as 93.4 estimated relevant documents in the estimation formulas (because it was selected for judging with probability 1/93.4). The otL07fbe run retrieved this document at rank 515; any other run that retrieved the document did so at the same or deeper rank (i.e., if the pooling had been to less than depth 515, the document would not have been in the pool). Note that the document content and metadata can be found online by appending the document identifier to the url "http://legacy.library.ucsf.edu/tid/" (e.g., `http://legacy.library.ucsf.edu/tid/hdz83f00`). One can do failure analysis for the final Boolean query for most topics by reviewing the underlined document identifiers.

Topic 52 (2007-A-1)

Request Text: Please produce any and all documents that discuss the use or introduction of high-phosphate fertilizers (HPF) for the specific purpose of boosting crop yield in commercial agriculture.

Initial Proposal by Defendant: `"high-phosphate fertilizer!" AND (boost! w/5 "crop yield") AND (commercial w/5 agricultur!)`

Rejoinder by Plaintiff: `(phosphat! OR hpf OR phosphorus OR fertiliz!) AND (yield! OR output OR produc! OR crop OR crops)`

Final Negotiated Boolean Query: `(("high-phosphat! fertiliz!" OR hpf) OR ((phosphat! OR phosphorus) w/15 (fertiliz! OR soil))) AND (boost! OR increas! OR rais! OR augment! OR affect! OR effect! OR multipl! OR doubl! OR tripl! OR high! OR greater) AND (yield! OR output OR produc! OR crop OR crops)`

Sampling: 361264 pooled, 1000 assessed, 55 judged relevant, 941 non-relevant, 4 gray, "C"=4.68, **Est. Rel.**: 257.4

Final Boolean Result Size (B): 3078, **Est. Recall**: 97.6%, **Est. Precision**: 10.6%

Participant High Recall@B: 100.0% (wat5nofeed), **Median Recall@B**: 48.2%

Participant High Recall@25000: 100.0% (wat1fuse and 10 others), **Median Recall@25000**: 73.1%

5 Deepest Sampled Relevant Documents: hdz83f00-93.4 (otL07fbe-515), dud53d00-21.8 (UMass15-106), huw23d00-18.8 (UMKC2-91), zge78d00-17.0 (SabL07ar1-82), ehe58c00-13.4 (wat5nofeed-64)

Topic 53 (2007-A-2)

Request Text: Please produce any and all documents concerning the effect of Maleic hydrazide (MH) on the tumorigenicity in hamsters.

Initial Proposal by Defendant: "maleic hydrazide" AND tumorigenicity AND hamster!

Rejoinder by Plaintiff: ("maleic hydr?zide" OR MH OR pesticid! OR "weed killer" OR herbicid! OR ((growth OR sprout!) w/3 (inhibitor! OR retardant)) OR "potassium salt" OR De-cut OR "Drexel MH" OR Gro-taro OR C4N2H4O2) AND (hamster! OR mice OR mouse OR rat OR rats OR rodent! OR subject! OR animal!)

Final Negotiated Boolean Query: ("maleic hydr?zide" OR (MH AND (pesticid! OR "weed killer" OR herbicid! OR (growth OR sprout!) w/3 (inhibitor! OR retardant))) OR "potassium salt" OR De-cut OR "Drexel MH" OR Gro-taro OR C4N2H4O2) AND (tumor! OR oncogenic OR oncology! OR pathology! OR pathogen!) AND (hamster! OR mice OR mouse OR rat OR rats OR rodent!)

Sampling: 309106 pooled, 499 assessed, 140 judged relevant, 359 non-relevant, 0 gray, “C”=2.26, **Est. Rel.**: 31632.5

Final Boolean Result Size (B): 4066, **Est. Recall**: 8.3%, **Est. Precision**: 60.7%

Participant High Recall@B: 12.0% (UMKC6), **Median Recall@B**: 3.0%

Participant High Recall@25000: 44.7% (ursinus4), **Median Recall@25000**: 16.7%

5 Deepest Sampled Relevant Documents: piq79c00-3407.2 (otL07rvl-24173), cdn65e00-3009.0 (UIowa07LegE5-17078), bin58d00-2556.7 (otL07fbe-11824), urb90d00-2284.6 (wat6qap-9507), pcp77c00-2194.5 (UMKC2-8839)

Topic 55 (2007-A-4)

Request Text: Please produce any and all documents concerning the known radioactivity of apatite rock.

Initial Proposal by Defendant: apatite w/15 radioactiv!

Rejoinder by Plaintiff: (Apatite OR "CA5(PO4)3OH" OR "CA5(PO4)3F" OR "CA5(PO4)3Cl" OR Fluorapatite OR Chlorapatite OR Hydroxylapatite) OR ((rock OR geolo!) AND (radioactiv! OR unstable OR instabil! OR radiat! OR radium OR polonium OR lead))

Final Negotiated Boolean Query: (Radioactiv! OR unstable OR instabil! OR radiat! OR radium OR polonium OR lead) AND (apatite OR "CA5(PO4)3OH" OR "CA5(PO4)3F" OR "CA5(PO4)3Cl" OR Fluorapatite OR Chlorapatite OR Hydroxylapatite)

Sampling: 380213 pooled, 496 assessed, 46 judged relevant, 440 non-relevant, 10 gray, “C”=1.27, **Est. Rel.**: 5564.7

Final Boolean Result Size (B): 580, **Est. Recall**: 1.7%, **Est. Precision**: 22.5%

Participant High Recall@B: 6.4% (wat4feed), **Median Recall@B**: 0.6%

Participant High Recall@25000: 52.2% (fdwim7ss), **Median Recall@25000**: 3.8%

5 Deepest Sampled Relevant Documents: vwc40e00-1871.6 (wat4feed-3799), dmm74e00-1507.2 (fdwim7ss-2740), jwr99d00-1289.2 (fdwim7ss-2206), qsj72f00-92.3 (UMKC1-574), dtn01e00-91.4 (wat4feed-548)

Topic 56 (2007-A-5)

Request Text: Please produce any and all documents concerning soil water management as it pertains to commercial irrigation.

Initial Proposal by Defendant: ("soil water" w/10 manage!) AND "commercial irrigation"

Rejoinder by Plaintiff: (Soil! OR sewage OR sewer! OR septic OR drain! OR dirt OR field! OR groundwater OR (ground w/3 water)) AND (manage! OR control!) AND irrigat!

Final Negotiated Boolean Query: (((Soil! OR sewage OR sewer! OR septic OR drain! OR dirt OR field! OR groundwater OR (ground w/3 water)) AND (manage! OR "control system")) AND irrigat!)

Sampling: 319017 pooled, 499 assessed, 112 judged relevant, 361 non-relevant, 26 gray, “C”=2.02, **Est. Rel.**: 2461.0

Final Boolean Result Size (B): 3288, **Est. Recall**: 46.8%, **Est. Precision**: 42.0%

Participant High Recall@B: 64.5% (UMKC5), **Median Recall@B**: 26.0%

Participant High Recall@25000: 87.3% (otL07frw), **Median Recall@25000**: 56.3%

5 Deepest Sampled Relevant Documents: nxw38d00-445.2 (ursinus8-2785), lmz34c00-421.3 (fdwim7ts-2369), amx11c00-291.1 (UMKC6-1055), rin34d00-283.6 (wat3desc-1007), cyp41a00-255.1 (UMKC2-842)

Topic 57 (2007-A-6)

Request Text: Please produce any and all documents that discuss methods for decreasing sugar loss in sugar-beet crops.

Initial Proposal by Defendant: "sugar beet" AND "sugar loss"

Rejoinder by Plaintiff: sugar! AND (beet OR beets OR crop OR crops) AND (lost OR loss OR losses OR decreas! OR wane! OR reduc! OR prevent!)

Final Negotiated Boolean Query: (sugar-beet OR sugarbeet OR beet OR beets OR crop OR crops) w/75 (lost OR loss OR losses OR decreas! OR wane! OR reduc! OR prevent!) w/75 sugar!

Sampling: 307648 pooled, 1000 assessed, 340 judged relevant, 643 non-relevant, 17 gray, “C”=4.67, **Est. Rel.**: 49048.1

Final Boolean Result Size (B): 3006, **Est. Recall**: 3.2%, **Est. Precision**: 64.4%

Participant High Recall@B: 5.7% (ursinus6), **Median Recall@B**: 2.9%

Participant High Recall@25000: 39.0% (wat4feed), **Median Recall@25000**: 13.4%

5 Deepest Sampled Relevant Documents: vet80a00-2564.3 (wat4feed-24583), urp52d00-2502.5 (wat4feed-23396), vxr81a00-2436.5 (fdwim7ss-22194), rrx98e00-2423.5 (catchup0701p-21963), fgo02a00-2354.2 (wat4feed-20777)

Topic 58 (2007-A-7)

Request Text: Please produce any and all documents that discuss health problems caused by HPF, including, but not limited to immune disorders, toxic myopathy, chronic fatigue syndrome, liver dysfunctions, irregular heart-beat, reactive depression, and memory loss.

Initial Proposal by Defendant: phosphat! AND ("immune disorder!" OR "toxic myopathy" OR "chronic fatigue syndrome" OR "liver dysfunction!" OR "irregular heart-beat" OR "reactive depression" OR "memory loss") AND (cause OR relate OR associate! OR derive! OR correlate!)

Rejoinder by Plaintiff: `(HPF OR phosphat! OR phosphorus OR fertiliz!) AND (illness! OR health OR disorder! OR toxic! OR "chronic fatigue" OR dysfunction! OR irregular OR memor! OR immun! OR myopath! OR liver! OR kidney! OR heart! OR depress! OR loss OR lost))`

Final Negotiated Boolean Query: `Phosphat! w/75 (caus! OR relat! OR assoc! OR derive! OR correlat!) w/75 (health OR disorder! OR toxic! OR "chronic fatigue" OR dysfunction! OR irregular OR memor! OR immun! OR myopath! OR liver! OR kidney! OR heart! OR depress! OR loss OR lost)`

Sampling: 346836 pooled, 495 assessed, 41 judged relevant, 454 non-relevant, 0 gray, "C"=1.37, **Est. Rel.**: 1150.6

Final Boolean Result Size (B): 8183, **Est. Recall**: 94.0%, **Est. Precision**: 11.8%

Participant High Recall@B: 94.0% (wat7bool and 1 other), **Median Recall@B**: 7.0%

Participant High Recall@25000: 94.9% (otL07fb2x), **Median Recall@25000**: 9.4%

5 Deepest Sampled Relevant Documents: rmw20d00-571.8 (otL07fb-1204), mqo61a00-284.1 (wat6qap-471), riy94d00-70.5 (wat8gram-101), bcx01d00-66.5 (wat8gram-95), zyd58c00-23.7 (wat4feed-33)

Topic 59 (2007-A-8)

Request Text: Please produce any and all studies, reports, discussions or analyses of the limestone quicklime wastewater treatment method that discusses this treatment's effectiveness in minimizing water contamination.

Initial Proposal by Defendant: `(limestone OR quicklime) AND "wastewater treatment"`

Rejoinder by Plaintiff: `(Limestone OR quicklime) AND (wastewater OR waste! OR water! OR sewage OR sewer! OR dispos! OR irrigate! OR well OR wells OR treat! OR purify! OR purification OR reduc! OR septic OR clean! OR steril! OR minim!)`

Final Negotiated Boolean Query: `(Limestone OR quicklime) w/75 (wastewater OR waste! OR water! OR sewage OR sewer! OR dispos! OR irrigate! OR well OR wells) w/75 (treat! OR purify! OR purification OR reduc! OR septic OR clean! OR steril! OR minim!)`

Sampling: 329585 pooled, 499 assessed, 15 judged relevant, 482 non-relevant, 2 gray, "C"=1.09, **Est. Rel.**: 111.8

Final Boolean Result Size (B): 240, **Est. Recall**: 0.9%, **Est. Precision**: 0.8%

Participant High Recall@B: 60.8% (UMKC4), **Median Recall@B**: 7.2%

Participant High Recall@25000: 100.0% (CMUL07ibp and 12 others), **Median Recall@25000**: 63.7%

5 Deepest Sampled Relevant Documents: yuz34f00-38.1 (UMKC4-202), uql43d00-29.6 (CMUL07ibp-84), aqo33d00-13.3 (CMUL07std-20), bti91f00-11.2 (CMUL07o1-16), eex53e00-9.6 (SabL07ar1-13)

Topic 60 (2007-A-9)

Request Text: Please produce any and all documents that discuss phosphate precipitation as a method of water purification.

Initial Proposal by Defendant: `(phosphate w/3 precipitation) AND (water w/3 purification)`

Rejoinder by Plaintiff: `phosphat! AND (precip! OR septic OR method!) AND purif!`

Final Negotiated Boolean Query: `(phosphat! w/75 (precip! OR septic OR method!)) AND ((water! OR waste!) w/75 purif!)`

Sampling: 279129 pooled, 700 assessed, 10 judged relevant, 669 non-relevant, 21 gray, "C"=2.49, **Est. Rel.**: 83.2

Final Boolean Result Size (B): 1496, **Est. Recall**: 7.2%, **Est. Precision**: 0.5%

Participant High Recall@B: 71.8% (UMass11 and 1 other), **Median Recall@B**: 7.6%

Participant High Recall@25000: 100.0% (CMUL07ibp and 16 others), **Median Recall@25000**: 90.6%

5 Deepest Sampled Relevant Documents: ake51c00-53.7 (UMass10-163), tcg42d00-18.1 (ursinus3-48), rbc63d00-4.4 (ursinus1-11), oma59c00-1.0 (UMass13-3), bmc55c00-1.0 (otL07fbe-3)

Topic 61 (2007-A-10)

Request Text: Please produce any and all waste treatment schedules that discuss phosphate concentrations in water.

Initial Proposal by Defendant: `("waste treatment" w/3 schedule!) AND (phosphate w/5 water)`

Rejoinder by Plaintiff: `schedul! AND (phosphat! OR phosphor!) AND (water OR waste! OR runoff OR irrigat! OR drain! OR sewage OR sewer OR liquid!)`

Final Negotiated Boolean Query: `Treat! w/150 schedul! w/150 (phosphat! OR phosphor!) w/150 (water OR waste! OR runoff OR irrigat! OR drain! OR sewage OR sewer OR liquid!)`

Sampling: 252532 pooled, 507 assessed, 64 judged relevant, 440 non-relevant, 3 gray, "C"=1.11, **Est. Rel.**: 372.1

Final Boolean Result Size (B): 296, **Est. Recall**: 43.9%, **Est. Precision**: 50.1%

Participant High Recall@B: 57.0% (otL07frw), **Median Recall@B**: 8.0%

Participant High Recall@25000: 98.9% (otL07fv), **Median Recall@25000**: 78.7%

5 Deepest Sampled Relevant Documents: bzz83e00-37.8 (wat6qap-116), una63e00-34.4 (otL07fb2x-91), nwe73e00-34.1 (otL07fb2x-89), txe73e00-29.4 (wat3desc-65), hpo02a00-27.0 (fdwim7sl-55)

Topic 62 (2007-A-11) *(This topic's assessments were completed too late for inclusion in the official set of 43 topics.)*

Request Text: Please produce any and all documents relating to any and all press releases concerning water contamination related to irrigation.

Initial Proposal by Defendant: `"press release!" AND "water contamination" AND irrigation`

Rejoinder by Plaintiff: `("press release" OR "press releases" OR news! OR report! OR public! OR announce! OR noti!) AND (contamina! OR toxic! OR pollut!) AND (irrigat! OR water)`

Final Negotiated Boolean Query: `("press release" OR "press releases" OR news! OR report! OR public! OR` `announce! OR noti!) w/150 water w/100 (contamina! OR toxic! OR pollut!) w/150 irrigat!`

Sampling: 355008 pooled, 499 assessed, 25 judged relevant, 466 non-relevant, 8 gray, "C"=0.50, **Est. Rel.**: 271.0

Final Boolean Result Size (B): 354, **Est. Recall**: 1.1%, **Est. Precision**: 1.9%

Participant High Recall@B: 47.5% (otL07fbe), **Median Recall@B**: 1.1%

Participant High Recall@25000: 67.3% (otL07fbe), **Median Recall@25000**: 30.3%

5 Deepest Sampled Relevant Documents: ahq25c00-64.0 (otL07fbe-334), aov39e00-59.6 (otL07fbe-189), bei86c00-57.8 (CMUL07o1-157), cul10a00-35.2 (UMKC6-35), ufm82c00-13.1 (CMUL07ibp-8)

Topic 63 (2007-A-12)

Request Text: Please produce any and all documents that specifically discuss an exclusivity clause in a sugar contract.

Initial Proposal by Defendant: `"sugar contract" AND "exclusivity clause"`

Rejoinder by Plaintiff: `Sugar AND (contract! OR agreement! OR deal! OR exclusiv!)`

Final Negotiated Boolean Query: `(Sugar w/20 (contract! OR agreement! OR deal!)) AND exclusiv!`

Sampling: 341624 pooled, 506 assessed, 11 judged relevant, 479 non-relevant, 16 gray, "C"=0.79, **Est. Rel.**: 18.6

Final Boolean Result Size (B): 294, **Est. Recall**: 26.8%, **Est. Precision**: 2.4%

Participant High Recall@B: 89.3% (UMass13), **Median Recall@B**: 16.1%

Participant High Recall@25000: 100.0% (otL07fv and 7 others), **Median Recall@25000**: 83.9%

5 Deepest Sampled Relevant Documents: hko97e00-8.6 (otL07pb-8), whv93d00-1.0 (wat1fuse-5), avt49c00-1.0 (otL07fv-4), gko97e00-1.0 (otL07pb-3), axb56e00-1.0 (UIowa07LegE1-3)

Topic 64 (2007-A-13)

Request Text: Please produce any and all documents that specifically discuss any and all deceptive implied health claims related to sugar.

Initial Proposal by Defendant: `deceptive AND (health w/5 claim!) w/75 sugar`

Rejoinder by Plaintiff: `(Decept! OR deceive OR false OR inaccurate OR mislead! OR misinform! OR misguid! OR untrue OR claim! OR state! OR declar! OR inform!) AND sugar`

Final Negotiated Boolean Query: `(Decept! OR deceive OR false OR inaccurate OR mislead! OR misinform! OR misguid! OR untrue) w/15 (claim! OR state OR statement! OR declare! OR inform!) w/75 sugar`

Sampling: 335933 pooled, 594 assessed, 16 judged relevant, 558 non-relevant, 20 gray, "C"=1.64, **Est. Rel.**: 159.1

Final Boolean Result Size (B): 131, **Est. Recall**: 6.9%, **Est. Precision**: 8.3%

Participant High Recall@B: 41.0% (wat4feed), **Median Recall@B**: 0.6%

Participant High Recall@25000: 99.4% (wat1fuse), **Median Recall@25000**: 3.8%

5 Deepest Sampled Relevant Documents: fzw07d00-79.8 (wat4feed-133), ubd22d00-18.2 (wat4feed-97), bbb63e00-14.1 (wat4feed-50), bhe58c00-13.1 (wat4feed-43), hvh77e00-11.9 (wat4feed-36)

Topic 65 (2007-A-14)

Request Text: Please produce any and all documents that explicitly discuss candy packaging, the labeling of candy, or which provide examples of candy packages or wrappers.

Initial Proposal by Defendant: `candy w/5 (packag! OR label! OR wrapper!)`

Rejoinder by Plaintiff: `Candy AND (pack! OR label! OR wrap! OR adverti! OR box OR ingredient! OR contain!)`

Final Negotiated Boolean Query: `Candy w/15 (pack! OR label! OR wrap! OR adverti! OR box OR ingredient! OR contain!)`

Sampling: 338958 pooled, 500 assessed, 58 judged relevant, 425 non-relevant, 17 gray, "C"=1.04, **Est. Rel.**: 609.7

Final Boolean Result Size (B): 8700, **Est. Recall**: 67.2%, **Est. Precision**: 7.8%

Participant High Recall@B: 97.0% (otL07rvl), **Median Recall@B**: 65.7%

Participant High Recall@25000: 99.2% (wat1fuse), **Median Recall@25000**: 68.8%

5 Deepest Sampled Relevant Documents: abh6aa00-159.8 (CMUL07ibp-183), aue3aa00-143.0 (SabL07ab1-162), czp15d00-75.4 (wat5nofeed-82), gtt92d00-41.3 (otL07fbe-44), rim00e00-35.8 (UMKC6-38)

Topic 66 (2007-A-15)

Request Text: Please produce any and all documents concerning the formation of the U.S. Beet Sugar Association.

Initial Proposal by Defendant: `((U.S. OR US) w/3 "Beet Sugar Association") AND (impact OR policy OR policies OR legislation)`

Rejoinder by Plaintiff: `("beet sugar" w/30 association!) AND (form OR formed OR start! OR create! OR founded OR began OR first OR initiat! OR initial OR begin! OR conceive!)`

Final Negotiated Boolean Query: `"beet sugar" AND association! AND (form OR formed OR start! OR create! OR founded OR began OR first OR initiat! OR initial OR begin! OR conceive!)`

Sampling: 415787 pooled, 488 assessed, 25 judged relevant, 455 non-relevant, 8 gray, "C"=1.01, **Est. Rel.**: 454.6

Final Boolean Result Size (B): 162, **Est. Recall**: 9.0%, **Est. Precision**: 27.1%

Participant High Recall@B: 9.0% (otL07fb2x and 1 other), **Median Recall@B**: 1.8%

Participant High Recall@25000: 99.1% (UMass11), **Median Recall@25000**: 6.7%

5 Deepest Sampled Relevant Documents: hls25e00-400.7 (UMass10-440), jgu96d00-21.4 (CMUL07ibt-64), ewn59e00-6.4 (wat6qap-8), ogi57d00-5.0 (wat7bool-6), apk48c00-1.0 (otL07pb-5)

Topic 67 (2007-A-16)

Request Text: Please produce any and all documents that explicitly refer to "The Sugar Program," and/or discuss the formation, contemplation or existence of a sugar cartel, or that discuss the sugar lobby in the context of Sugar Acts passed by Congress.

Initial Proposal by Defendant: "Sugar Program" AND "sugar cartel"

Rejoinder by Plaintiff: "Sugar Program" OR (Sugar AND (lobby! OR Congress OR "sugar acts" OR law! OR polic! OR legis! OR regulat! OR ordinance! OR control! OR cartel! OR combine OR syndicate OR trust OR conspir!))

Final Negotiated Boolean Query: "Sugar Program" OR ((sugar OR sucrose) AND (cartel OR combine)) OR ((Sugar OR sucrose) w/15 (lobby! OR Congress OR "sugar acts" OR law! OR polic! OR legis! OR regulat! OR ordinance! OR control!))

Sampling: 383624 pooled, 493 assessed, 75 judged relevant, 418 non-relevant, 0 gray, "C"=1.01, **Est. Rel.**: 41189.6

Final Boolean Result Size (B): 13241, **Est. Recall**: 1.4%, **Est. Precision**: 5.7%

Participant High Recall@B: 13.0% (UMass12), **Median Recall@B**: 3.8%

Participant High Recall@25000: 26.9% (ursinus8), **Median Recall@25000**: 8.0%

5 Deepest Sampled Relevant Documents: lgo18d00-4085.5 (ursinus8-22561), idj24d00-3706.9 (UIowa07LegE3-14476), gog85a00-3670.2 (UMKC3-13938), izo81d00-3627.2 (SabL07ar2-13343), clc07e00-2190.5 (UIowa07LegE0-12801)

Topic 69 (2007-B-1)

Request Text: All documents referring or relating to indoor smoke ventilation.

Initial Proposal by Defendant: "indoor smoke ventilation"

Rejoinder by Plaintiff: (indoor OR inside) AND ("smoke ventilation" OR filtration)

Final Negotiated Boolean Query: (indoor OR inside) w/25 ("smoke ventilation" OR filtration)

Sampling: 226267 pooled, 495 assessed, 280 judged relevant, 110 non-relevant, 105 gray, "C"=1.55, **Est. Rel.**: 37457.5

Final Boolean Result Size (B): 5123, **Est. Recall**: 8.3%, **Est. Precision**: 96.9%

Participant High Recall@B: 13.7% (wat2nobool), **Median Recall@B**: 11.5%

Participant High Recall@25000: 61.0% (UMKC3), **Median Recall@25000**: 39.6%

5 Deepest Sampled Relevant Documents: khk42d00-3732.5 (catchup0701p-22822), vyh91f00-2931.7 (SabL07ar1-10985), bfm91f00-2212.0 (UMKC3-6149), mjf08d00-2122.2 (otL07fbe-5715), art52c00-2042.9 (CMUL07ibs-5354)

Topic 70 (2007-B-2)

Request Text: All documents that make reference to the smell of baked goods, including but not limited to baked cookies.

Initial Proposal by Defendant: smell w/3 ("baked goods" OR "baked cookie!")

Rejoinder by Plaintiff: (smell OR aroma) AND baked OR pie! OR bread! OR cake OR food!

Final Negotiated Boolean Query: (smell OR aroma) w/15 ("baked good!" OR "baked cookie!" OR pie! OR bread! OR cake! OR foodstuff!)

Sampling: 352690 pooled, 499 assessed, 43 judged relevant, 452 non-relevant, 4 gray, "C"=1.26, **Est. Rel.**: 19828.1

Final Boolean Result Size (B): 1381, **Est. Recall**: 0.1%, **Est. Precision**: 1.7%

Participant High Recall@B: 1.1% (UIowa07LegE2), **Median Recall@B**: 0.1%

Participant High Recall@25000: 35.4% (ursinus4), **Median Recall@25000**: 1.6%

5 Deepest Sampled Relevant Documents: ajk32d00-3551.3 (ursinus4-15444), eth06c00-2631.9 (UIowa07LegE1-7002), bue80c00-2388.1 (otL07fbe-5760), fyl99d00-2202.2 (UIowa07LegE1-4959), eli23e00-1957.4 (ursinus1-4053)

Topic 71 (2007-B-3)

Request Text: All documents discussing the condition of bromhidrosis (a/k/a body odor).

Initial Proposal by Defendant: bromhidrosis

Rejoinder by Plaintiff: bromhidrosis OR ((body OR human OR person) AND odor!))

Final Negotiated Boolean Query: bromhidrosis OR ((body OR human OR person) w/3 odor!)

Sampling: 356441 pooled, 697 assessed, 308 judged relevant, 376 non-relevant, 13 gray, "C"=2.28, **Est. Rel.**: 77466.9

Final Boolean Result Size (B): 4527, **Est. Recall**: 4.3%, **Est. Precision**: 69.5%

Participant High Recall@B: 5.8% (CMUL07ibt and 2 others), **Median Recall@B**: 3.4%

Participant High Recall@25000: 23.1% (otL07pb), **Median Recall@25000**: 11.2%

5 Deepest Sampled Relevant Documents: myo01d00-3186.1 (wat4feed-20024), rsi55d00-3112.8 (ursinus5-18803), cqr42e00-3084.7 (fdwim7xj-18361), tfp94a00-3049.7 (IowaSL07Ref-17826), dns25e00-2710.8 (UMKC5-13499)

Topic 72 (2007-B-4)

Request Text: All documents referring to the scientific or chemical process(es) which result in onions have the effect of making persons cry.

Initial Proposal by Defendant: ("scientific process! OR "chemical process!") AND onion AND cry

Rejoinder by Plaintiff: ((scien! OR research! OR chemical) AND onion!) AND (cries OR cry! OR tear!)

Final Negotiated Boolean Query: ((scien! OR research! OR chemical) w/25 onion!) AND (cries OR cry! OR tear!)

Sampling: 400625 pooled, 499 assessed, 11 judged relevant, 477 non-relevant, 11 gray, "C"=0.60, **Est. Rel.**: 97.7

Final Boolean Result Size (B): 119, **Est. Recall**: 77.9%, **Est. Precision**: 43.0%

Participant High Recall@B: 77.9% (otL07fb), **Median Recall@B**: 3.1%

Participant High Recall@25000: 100.0% (UMKC5 and 6 others), **Median Recall@25000**: 50.5%

5 Deepest Sampled Relevant Documents: tmj97c00-20.7 (otL07fb-94), hes71f00-20.6 (otL07fbe-91), brq10e00-18.8

(wat7bool-54), cmj97c00-12.6 (otL07frw-16), vlj97c00-12.2 (CMUL07ibp-15)

Topic 73 (2007-B-5)

Request Text: Any advertisements or draft advertisements that target women seen in the kitchen or cooking.

Initial Proposal by Defendant: `advertisement AND "target! women"`

Rejoinder by Plaintiff: `(("ad campaign" OR advertis!) AND (woman OR women OR girl OR female) AND (kitchen OR cook!)`

Final Negotiated Boolean Query: `(("ad campaign" OR advertis!) w/25 (woman OR women OR girl OR female)) AND (kitchen OR cook!)`

Sampling: 370452 pooled, 498 assessed, 72 judged relevant, 426 non-relevant, 0 gray, "C"=0.74, **Est. Rel.**: 31894.5

Final Boolean Result Size (B): 4085, **Est. Recall**: 4.9%, **Est. Precision**: 41.5%

Participant High Recall@B: 8.7% (UMKC5), **Median Recall@B**: 1.0%

Participant High Recall@25000: 51.6% (UMKC2), **Median Recall@25000**: 4.9%

5 Deepest Sampled Relevant Documents: tdc58e00-4345.7 (UMKC2-24574), qyf46e00-4279.2 (UIowa07LegE5-21967), nth79d00-4259.8 (CMUL07irs-21294), myc81a00-4067.3 (UIowa07LegE5-16135), djv94c00-3504.2 (wat6qap-8668)

Topic 74 (2007-B-6)

Request Text: All scientific studies expressly referencing health effects tied to indoor air quality.

Initial Proposal by Defendant: `"health effect!" w/10 "air quality"`

Rejoinder by Plaintiff: `(scien! OR stud! OR research) AND ("air quality" OR health)`

Final Negotiated Boolean Query: `(scien! OR stud! OR research) AND ("air quality" w/15 health)`

Sampling: 225883 pooled, 499 assessed, 299 judged relevant, 196 non-relevant, 4 gray, "C"=1.50, **Est. Rel.**: 62406.2

Final Boolean Result Size (B): 20516, **Est. Recall**: 22.2%, **Est. Precision**: 77.2%

Participant High Recall@B: 32.8% (wat3desc), **Median Recall@B**: 24.3%

Participant High Recall@25000: 40.0% (UMKC4), **Median Recall@25000**: 29.2%

5 Deepest Sampled Relevant Documents: vkm92d00-3123.0 (Dartmouth1-19609), gew24e00-3095.9 (SabL07ab1-18917), ljq87c00-3070.3 (otL07fbe-18296), mhf57d00-2942.7 (UIowa07LegE3-15606), bof25d00-2912.1 (otL07pb-15048)

Topic 75 (2007-B-7)

Request Text: All documents that memorialize any statement or suggestion from an elected federal public official that further research is necessary to improve indoor environmental air quality.

Initial Proposal by Defendant: `"public official" AND research w/10 ("indoor air quality" OR "indoor environment! air quality")`

Rejoinder by Plaintiff: `(("public official" OR senator OR representative OR congressman OR congresswoman OR president OR vice-president OR VP) AND ((research OR scienc! OR stud!) AND (indoor AND (environment! w/5 "air quality") AND (statement OR "public debate" OR suggestion OR remark!)`

Final Negotiated Boolean Query: `("public official" OR senator OR representative OR congressman OR congresswoman OR president OR vice-president OR VP) AND ((research OR scienc! OR stud!) w/25 indoor w/25 environment! w/5 "air quality") AND (statement OR "public debate" OR suggestion OR remark!)`

Sampling: 263784 pooled, 499 assessed, 23 judged relevant, 469 non-relevant, 7 gray, "C"=0.91, **Est. Rel.**: 228.1

Final Boolean Result Size (B): 788, **Est. Recall**: 13.5%, **Est. Precision**: 5.3%

Participant High Recall@B: 45.6% (SabL07ab1 and 1 other), **Median Recall@B**: 4.0%

Participant High Recall@25000: 100.0% (fdwim7ss and 11 others), **Median Recall@25000**: 86.1%

5 Deepest Sampled Relevant Documents: bxc52c00-89.4 (UMKC4-188), zku96d00-60.8 (SabL07ab1-90), kiu34e00-30.2 (SabL07ab1-34), twh67c00-28.7 (otL07fb2x-32), iif12f00-1.0 (CMUL07o1-5)

Topic 76 (2007-B-8)

Request Text: All documents that make reference to any public meeting or conference held in Washington, D.C. on the subject of indoor air quality.

Initial Proposal by Defendant: `("public meeting" OR conference) AND "Washington, D.C." AND "indoor air quality"`

Rejoinder by Plaintiff: `((meeting OR conference OR event OR symposium) AND (Washington! OR "D.C." OR "District of Columbia") AND (indoor AND "air quality")`

Final Negotiated Boolean Query: `(meeting OR conference OR event OR symposium) AND (Washington! OR "D.C." OR "District of Columbia") AND (indoor w/5 "air quality")`

Sampling: 195688 pooled, 1000 assessed, 254 judged relevant, 730 non-relevant, 16 gray, "C"=6.49, **Est. Rel.**: 4408.9

Final Boolean Result Size (B): 22518, **Est. Recall**: 50.8%, **Est. Precision**: 10.9%

Participant High Recall@B: 77.8% (CMUL07std), **Median Recall@B**: 50.8%

Participant High Recall@25000: 78.4% (ursinus2), **Median Recall@25000**: 52.1%

5 Deepest Sampled Relevant Documents: ijz83e00-415.2 (UIowa07LegE3-2968), qwf00a00-403.1 (UIowa07LegE1-2873), agq30c00-396.2 (UIowa07LegE3-2819), bwf69d00-345.7 (fdwim7ts-2430), jxp57d00-304.8 (SabL07ar1-2122)

Topic 77 (2007-B-9)

Request Text: All documents that refer or relate to the effect of smoke on bystanders, excluding studies on cigarette smoking or tobacco smoke.

Initial Proposal by Defendant: `(effect AND smoke) w/5 bystander BUT NOT (tobacco OR cigarette)`

Rejoinder by Plaintiff: `(smok! OR effect) AND (bystander! OR "third party" OR passive!)`

Final Negotiated Boolean Query: `(smok! AND bystander!) BUT NOT (tobacco OR cigarette)`

Sampling: 345347 pooled, 499 assessed, 11 judged relevant, 477 non-relevant, 11 gray, "C"=0.83, **Est. Rel.**: 11233.8

Final Boolean Result Size (B): 154, **Est. Recall**: 0.0%, **Est. Precision**: 0.0%

Participant High Recall@B: 0.2% (wat4feed), **Median Recall@B**: 0.0%

Participant High Recall@25000: 53.1% (wat4feed), **Median Recall@25000**: 0.0%

5 Deepest Sampled Relevant Documents: ivm59c00-4116.8 (otL07rvl-19345), zul52d00-3883.9 (wat4feed-14441), usn97c00-1805.7 (wat4feed-2346), gof44c00-1153.1 (IowaSL0706-1244), rrp87d00-250.6 (wat4feed-219)

Topic 78 (2007-B-10)

Request Text: All documents referencing patents on odors, excluding tobacco or cigarette related patents.

Initial Proposal by Defendant: `(patent! w/15 odor!) BUT NOT (tobacco OR cigarette)`

Rejoinder by Plaintiff: `patent! AND odor!`

Final Negotiated Boolean Query: `(patent! AND odor!) BUT NOT (tobacco OR cigarette!)`

Sampling: 288711 pooled, 499 assessed, 24 judged relevant, 440 non-relevant, 35 gray, "C"=1.30, **Est. Rel.**: 835.4

Final Boolean Result Size (B): 1611, **Est. Recall**: 25.3%, **Est. Precision**: 11.7%

Participant High Recall@B: 61.3% (otL07pb), **Median Recall@B**: 7.7%

Participant High Recall@25000: 99.9% (CMUL07ibt and 7 others), **Median Recall@25000**: 38.4%

5 Deepest Sampled Relevant Documents: anc13c00-232.2 (wat8gram-1081), agu15d00-191.2 (IowaSL0704-611), xph02a00-130.5 (UIowa07LegE2-285), wau60c00-52.2 (otL07fv-81), gnb97c00-45.0 (otL07pb-68)

Topic 79 (2007-C-1)

Request Text: All documents making a connection between the music and songs of Peter, Paul, and Mary, Joan Baez, or Bob Dylan, and the sale of cigarettes.

Initial Proposal by Defendant: `(music OR songs) AND (((peter w/2 paul AND (paul w/2 mary)) OR "bob Dylan" OR "joan baez") AND sale`

Rejoinder by Plaintiff: `((peter AND paul AND mary) OR dylan OR baez) AND (sale! OR sell! OR advertis! OR promot! OR market!)`

Final Negotiated Boolean Query: `(((peter w/3 paul) AND (paul w/3 mary)) OR (simon w/3 garfunkel) OR "Bob Dylan" OR "Joan Baez") AND (sale! OR sell! OR advertis! OR promot! OR market!)`

Sampling: 409225 pooled, 491 assessed, 35 judged relevant, 448 non-relevant, 8 gray, "C"=1.03, **Est. Rel.**: 1486.6

Final Boolean Result Size (B): 317, **Est. Recall**: 6.5%, **Est. Precision**: 50.2%

Participant High Recall@B: 9.5% (otL07frw), **Median Recall@B**: 3.3%

Participant High Recall@25000: 100.0% (otL07fbe), **Median Recall@25000**: 10.9%

5 Deepest Sampled Relevant Documents: thr11a00-766.2 (otL07fbe-932), ecp25d00-478.5 (IowaSL0706-545), bea05d00-51.9 (otL07fbe-294), goq26e00-48.3 (SabL07arbn-208), dhv36c00-28.4 (fdwim7xj-53)

Topic 80 (2007-C-2)

Request Text: All documents making a connection between folk songs and music and the sale of cigarettes.

Initial Proposal by Defendant: `"folk songs" AND (sale! OR sell! OR promot! OR advertis! OR market!)`

Rejoinder by Plaintiff: `folk AND (sale OR sell! OR promot! OR advertis! OR market!)`

Final Negotiated Boolean Query: `("folk songs" OR "folk music" OR "folk artists") AND (sale! OR sell! OR promot! OR advertis! OR market!)`

Sampling: 364619 pooled, 999 assessed, 391 judged relevant, 602 non-relevant, 6 gray, "C"=4.40, **Est. Rel.**: 38649.9

Final Boolean Result Size (B): 331, **Est. Recall**: 0.6%, **Est. Precision**: 81.9%

Participant High Recall@B: 0.8% (otL07rvl and 1 other), **Median Recall@B**: 0.6%

Participant High Recall@25000: 39.9% (otL07rvl and 1 other), **Median Recall@25000**: 15.8%

5 Deepest Sampled Relevant Documents: vxf58c00-2519.3 (UIowa07LegE0-22342), oxb28d00-2496.5 (ursinus5-21938), cwq61c00-2392.0 (IowaSL0702-20178), iyb70f00-2362.3 (UMKC6-19703), bcm43f00-2074.0 (wat4feed-15594)

Topic 82 (2007-C-4)

Request Text: All documents discussing the color of the paper used to make cigarettes in connection with increasing sales.

Initial Proposal by Defendant: `(color! w/2 paper) AND (increas! w/3 sales)`

Rejoinder by Plaintiff: `(color! OR shade! OR pastel! OR tint!) AND paper AND (sale! OR sell!)`

Final Negotiated Boolean Query: `((color! OR shade! OR pastel! OR tint!) w/5 paper) AND (increas! w/15 (sale! OR sell!))`

Sampling: 418281 pooled, 491 assessed, 176 judged relevant, 310 non-relevant, 5 gray, "C"=0.34, **Est. Rel.**: 75558.9

Final Boolean Result Size (B): 888, **Est. Recall**: 0.8%, **Est. Precision**: 61.2%

Participant High Recall@B: 1.1% (otL07pb), **Median Recall@B**: 0.4%

Participant High Recall@25000: 33.1% (otL07fbe), **Median Recall@25000**: 4.7%

5 Deepest Sampled Relevant Documents: nlq90a00-4664.5 (UMass10-23634), bim31f00-4627.1 (otL07fbe-21093), qwf74f00-4596.9 (otL07fbe-19384), dmv71d00-4396.0 (CMUL07irs-12373), hko20f00-4366.2 (otL07fbe-11712)

Topic 83 (2007-C-5)

Request Text: All documents discussing using psychedelic colors to increase sales of cigarettes.

Initial Proposal by Defendant: "psychedelic color!" AND (increas! w/3 sales)

Rejoinder by Plaintiff: psyched*lic AND (sale! OR sell! OR promot! OR advertis! OR market

Final Negotiated Boolean Query: psyched*lic AND color! AND (sale! OR sell! OR promot! OR advertis! OR market!)

Sampling: 385402 pooled, 496 assessed, 44 judged relevant, 452 non-relevant, 0 gray, "C"=0.52, **Est. Rel.**: 13987.5

Final Boolean Result Size (B): 281, **Est. Recall**: 0.8%, **Est. Precision**: 32.1%

Participant High Recall@B: 1.2% (otL07frw), **Median Recall@B**: 0.4%

Participant High Recall@25000: 33.3% (wat4feed), **Median Recall@25000**: 1.6%

5 Deepest Sampled Relevant Documents: bco01e00-4467.9 (wat4feed-21831), aqm49d00-4228.5 (UMass10-14250), xfk10d00-3935.5 (fdwim7sl-9612), ait55f00-967.7 (ursinus4-624), ect68c00-45.9 (SabL07ab1-131)

Topic 84 (2007-C-6)

Request Text: All documents referencing "Bonnie and Clyde" or a James Bond film or the films of Stanley Kubrick in connection with sales of cigarettes.

Initial Proposal by Defendant: (Bonnie w/2 Clyde) OR "James Bond film" OR "Stanley Kubrick"

Rejoinder by Plaintiff: ((Bonnie w/3 Clyde) OR ("James Bond" OR "Agent 007" OR Goldfinger OR "Dr. No" OR "Dr No" OR "From Russia With Love" OR Thunderball OR "Sean Connery" OR "Roger Moore") OR ("Stanley Kubrick" OR (Kubrick w/3 (movie! OR film)) OR "2001: A Space Odyssey" OR HAL OR Lolita OR "James Mason" OR "Dr Strangelove" OR "Dr. Strangelove" OR "Peter Sellers")) AND (sale! OR sell! OR promot! OR advertis! OR market!)

Final Negotiated Boolean Query: ((Bonnie w/3 Clyde) OR ("James Bond" OR "Agent 007" OR Goldfinger OR "Dr No" OR "Dr. No" OR "From Russia With Love" OR Thunderball) OR ("Stanley Kubrick" OR (Kubrick w/3 (movie! OR film)) OR "2001: A Space Odyssey" OR Lolita OR "Dr Strangelove" OR "Dr. Strangelove")) AND (sale! OR sell! OR promot! OR advertis! OR market!)

Sampling: 476252 pooled, 498 assessed, 63 judged relevant, 431 non-relevant, 4 gray, "C"=0.73, **Est. Rel.**: 450.1

Final Boolean Result Size (B): 2493, **Est. Recall**: 100.0%, **Est. Precision**: 18.6%

Participant High Recall@B: 100.0% (CMUL07ibp and 3 others), **Median Recall@B**: 38.3%

Participant High Recall@25000: 100.0% (CMUL07ibp and 7 others), **Median Recall@25000**: 85.7%

5 Deepest Sampled Relevant Documents: mrj70e00-137.8 (CMUL07ibs-139), qcm09c00-71.6 (otL07fb-61), gqd62d00-41.4 (wat5nofeed-33), mel03f00-41.4 (ursinus7-33), gsd19d00-29.6 (otL07db-23)

Topic 85 (2007-C-7)

Request Text: All documents discussing or referencing generally accepted accounting principles in connection with the decision to record as sales products shipped to distributors on a sale-or-return basis, and the implementation thereof.

Initial Proposal by Defendant: ("gaap" OR "generally accepted accounting principle!") AND (revenue! OR records OR recording OR account!)) AND (sale w/5 return)

Rejoinder by Plaintiff: ("gaap" OR "generally accepted accounting principle!" OR "fasb" OR "financial accounting standards board" OR "sab" OR "Staff accounting bulletin" OR "sas" OR (statement w/2 "auditing standards")) AND ((sale! OR allowance OR reserve! OR right! OR entitle! OR could) w/5 return!)

Final Negotiated Boolean Query: ("gaap" OR "generally accepted accounting principle!" OR "fasb" OR "financial accounting standards board" OR "sab" OR "Staff accounting bulletin" OR "sas" OR (statement w/2 "auditing standards")) AND (revenue! OR recording OR records OR account!) AND (sale! OR allowance OR reserve!) AND ((right! OR entitle! OR could) w/5 return!)

Sampling: 361317 pooled, 497 assessed, 96 judged relevant, 392 non-relevant, 9 gray, "C"=0.80, **Est. Rel.**: 3890.7

Final Boolean Result Size (B): 1305, **Est. Recall**: 13.8%, **Est. Precision**: 44.3%

Participant High Recall@B: 31.6% (IowaSL0705 and 1 other), **Median Recall@B**: 9.8%

Participant High Recall@25000: 77.5% (ursinus7), **Median Recall@25000**: 42.8%

5 Deepest Sampled Relevant Documents: lma14c00-221.5 (otL07fbe-1170), yip12d00-214.3 (ursinus3-957), ynz51d00-206.1 (SabL07ar1-783), yde76d00-203.7 (otL07fb-742), xsh35f00-201.7 (UIowa07LegE2-710)

Topic 86 (2007-C-8)

Request Text: All documents discussing or referencing both generally accepted accounting principles and the defendants' decision to restate its financial results.

Initial Proposal by Defendant: restate AND ("gaap" OR "generally accepted accounting principle!" OR "financial accounting standards board" OR "staff accounting bulletin" OR (statement w/2 "auditing standards")) AND (revenue! OR records OR recording)

Rejoinder by Plaintiff: (restate! OR revise!) AND (gaap OR "generally accepted accounting principle!" OR fasb OR financial OR sab OR accounting OR sas OR auditing

Final Negotiated Boolean Query: (restate! OR revise!) AND ("gaap" OR "generally accepted accounting principle!" OR "fasb" OR "financial accounting standards board" OR "sab" OR "staff accounting bulletin" OR "sas" OR (statement w/2 "auditing standards")) AND (revenue! OR records OR recording)

Sampling: 370179 pooled, 499 assessed, 21 judged relevant, 465 non-relevant, 13 gray, "C"=1.21, **Est. Rel.**: 8830.1

Final Boolean Result Size (B): 6446, **Est. Recall**: 4.8%, **Est. Precision**: 8.2%

Participant High Recall@B: 24.2% (UIowa07LegE0), **Median Recall@B**: 4.8%

Participant High Recall@25000: 51.2% (IowaSL07Ref), **Median Recall@25000**: 6.9%

5 Deepest Sampled Relevant Documents: gnl85a00-3408.1 (otL07fbe-12953), vgm85c00-981.3 (wat4feed-4971), ohk68c00-830.7 (wat3desc-2826), jbm58c00-755.8 (UIowa07LegE0-2210), rsc58c00-607.5 (catchup0701p-1390)

Topic 87 (2007-C-9)

Request Text: All documents discussing Securities and Exchange Commission 10b-5 reports or reporting requirements.

Initial Proposal by Defendant: "10b-5 Report!" AND (Securities w/3 "exchange commission")

Rejoinder by Plaintiff: 10! AND (SEC OR (Securities w/3 "Exchange Commission"))

Final Negotiated Boolean Query: 10b-5 AND (SEC OR (securities w/3 "exchange commission"))

Sampling: 351913 pooled, 496 assessed, 41 judged relevant, 444 non-relevant, 11 gray, "C"=1.36, **Est. Rel.**: 875.3

Final Boolean Result Size (B): 138, **Est. Recall**: 3.8%, **Est. Precision**: 40.1%

Participant High Recall@B: 6.5% (UMKC5 and 1 other), **Median Recall@B**: 1.8%

Participant High Recall@25000: 100.0% (UMKC2 and 6 others), **Median Recall@25000**: 11.0%

5 Deepest Sampled Relevant Documents: bgb65a00-758.0 (UMKC5-1215), hse51c00-21.5 (UMass12-132), ckq92f00-18.0 (fdwim7ts-70), xbn93c00-13.1 (UMKC1-34), cqp92e00-7.5 (wat8gram-14)

Topic 89 (2007-D-1)

Request Text: Submit all documents listing monthly and/or annual sales for companies in the property and casualty insurance business in the United States between 1980 and the present.

Initial Proposal by Defendant: (("monthly sales" OR "annual sales") AND ("property insurance" OR "casualty insurance")) AND (United States OR U.S.) AND (198* OR 199*)

Rejoinder by Plaintiff: (month! OR annual!) AND (sales OR sell! OR revenue) AND insurance AND (198* OR 199*)

Final Negotiated Boolean Query: (((month! OR annual!) w/15 (sales OR sell! OR revenue)) AND ((property OR casualty) AND insurance)) BUT NOT (England OR "Great Britain" OR U.K. OR UK)

Sampling: 345011 pooled, 496 assessed, 78 judged relevant, 392 non-relevant, 26 gray, "C"=1.16, **Est. Rel.**: 6083.6

Final Boolean Result Size (B): 3636, **Est. Recall**: 9.9%, **Est. Precision**: 16.8%

Participant High Recall@B: 21.0% (wat3desc), **Median Recall@B**: 8.9%

Participant High Recall@25000: 91.9% (otL07fbe), **Median Recall@25000**: 17.1%

5 Deepest Sampled Relevant Documents: mwv44d00-3850.5 (otL07fbe-19428), qxd35a00-561.1 (SabL07ab1-2850), qhc48c00-495.2 (IowaSL0707-1800), jyl13d00-259.1 (otL07fbe-467), ltq35f00-203.1 (SabL07ab1-327)

Topic 90 (2007-D-2)

Request Text: Submit all documents listing monthly and/or annual sales for companies in the property and casualty insurance business in England for all available years.

Initial Proposal by Defendant: (("monthly sales" OR "annual sales") AND ("property insurance" OR "casualty insurance")) AND England

Rejoinder by Plaintiff: (month! OR annual!) AND (sales OR sell! OR revenue) AND Insurance AND (England OR Brit! OR U.K. OR UK)

Final Negotiated Boolean Query: (((month! OR annual!) w/15 sales) AND ((property OR casualty) AND insurance)) AND (England OR "Great Britain" OR U.K. OR UK)

Sampling: 330155 pooled, 492 assessed, 34 judged relevant, 458 non-relevant, 0 gray, "C"=1.30, **Est. Rel.**: 1066.1

Final Boolean Result Size (B): 2665, **Est. Recall**: 10.3%, **Est. Precision**: 9.0%

Participant High Recall@B: 63.9% (otL07fbe), **Median Recall@B**: 14.6%

Participant High Recall@25000: 99.3% (otL07fbe), **Median Recall@25000**: 33.7%

5 Deepest Sampled Relevant Documents: fds95c00-393.5 (otL07fbe-1954), umo55a00-159.9 (CMUL07irt-297), wyu71f00-153.4 (wat4feed-280), cmr03f00-91.2 (otL07pb-143), hre33f00-85.8 (otL07fbe-133)

Topic 92 (2007-D-4)

Request Text: Submit all documents relating to competition or market share in the property and casualty insurance industry, including, but not limited to, market studies, forecasts and surveys.

Initial Proposal by Defendant: ("market stud!" OR forecast! OR survey!) AND "market share" AND ("property insurance" OR "casualty insurance")

Rejoinder by Plaintiff: (competition OR market OR share) AND insurance

Final Negotiated Boolean Query: ("market stud!" OR forecast! OR survey!) AND (competition OR share) AND (property OR casualty) AND insurance

Sampling: 313137 pooled, 498 assessed, 117 judged relevant, 369 non-relevant, 12 gray, "C"=1.63, **Est. Rel.**: 18070.2

Final Boolean Result Size (B): 9401, **Est. Recall**: 12.8%, **Est. Precision**: 21.0%

Participant High Recall@B: 18.6% (ursinus6), **Median Recall@B**: 8.6%

Participant High Recall@25000: 36.0% (UMass15), **Median Recall@25000**: 13.5%

5 Deepest Sampled Relevant Documents: ahe53c00-3532.2 (UMass13-19613), ggr75d00-3162.8 (wat8gram-14031), pkr48d00-2733.5 (CMUL07irt-9829), cxu05d00-1413.5 (SabL07ab1-9283), ams51d00-1309.8 (ursinus6-7038)

Topic 94 (2007-D-6)

Request Text: Submit all documents relating to insurance price lists, pricing plans, pricing policies, pricing forecasts, pricing strategies, pricing analyses, and pricing decisions.

Initial Proposal by Defendant: ("price lists" OR "pricing plans" OR "pricing policies" OR "pricing forecasts" OR "pricing strategies" OR "pricing analyses" OR "pricing decisions") AND ("property insurance" OR "casualty insurance")

Rejoinder by Plaintiff: ((price OR pricing) AND (list! OR plan! OR polic! OR Forecast! OR strateg! OR analys! OR decision!)) AND insurance

Final Negotiated Boolean Query: ((price OR pricing) w/15 (list! OR plan! OR polic! OR forecast! OR strateg! OR analys! OR decision!)) AND insurance

Sampling: 279484 pooled, 500 assessed, 104 judged relevant, 391 non-relevant, 5 gray, “C”=1.36, **Est. Rel.**: 40068.5

Final Boolean Result Size (B): 12080, **Est. Recall**: 6.7%, **Est. Precision**: 23.7%

Participant High Recall@B: 21.4% (CMUL07irs), **Median Recall@B**: 4.5%

Participant High Recall@25000: 45.5% (CMUL07irs), **Median Recall@25000**: 7.5%

5 Deepest Sampled Relevant Documents: nhw23f00-3901.8 (UMKC3-24159), fek48d00-3813.1 (CMUL07irs-21845), bsv84a00-3521.7 (wat6qap-16200), clg85a00-3281.1 (wat6qap-12980), ayy84a00-1831.4 (UIowa07LegE2-10294)

Topic 95 (2007-D-7)

Request Text: Submit all documents discussing or relating to the historical, current, or future financial impact of tobacco usage on the property and casualty insurance industry.

Initial Proposal by Defendant: ("historical" OR "current" OR "future") AND "financial impact" AND usage AND ("property insurance" OR "casualty insurance")

Rejoinder by Plaintiff: (financial OR (increas! w/3 cost!) OR (smoking w/5 (illness OR sick! OR death!)) AND insurance

Final Negotiated Boolean Query: ("financial impact" OR (increas! w/3 cost!) OR ("smoking-related" w/5 (illness OR sick! OR death!)) AND insurance

Sampling: 315430 pooled, 499 assessed, 120 judged relevant, 379 non-relevant, 0 gray, “C”=1.53, **Est. Rel.**: 34111.6

Final Boolean Result Size (B): 16324, **Est. Recall**: 18.5%, **Est. Precision**: 33.5%

Participant High Recall@B: 27.1% (CMUL07ibt), **Median Recall@B**: 8.7%

Participant High Recall@25000: 36.6% (CMUL07ibt), **Median Recall@25000**: 12.9%

5 Deepest Sampled Relevant Documents: qns51a00-3690.8 (otL07fbe-21566), gvn44a00-3435.7 (CMUL07irs-16802), pem65a00-2424.3 (UMKC1-14407), yns76d00-2418.1 (CMUL07ibp-14266), yxb15a00-2363.2 (UMass10-13092)

Topic 96 (2007-D-8)

Request Text: Submit all documents that discuss entry conditions into the property and casualty insurance industry.

Initial Proposal by Defendant: "entry condition!" AND ("property insurance" OR "casualty insurance"

Rejoinder by Plaintiff: (entry AND (barrier! OR condition!)) AND ((property OR casualty) w/10 insurance)

Final Negotiated Boolean Query: (entry w/10 (barrier! OR condition!)) AND ((property OR casualty) w/10 insurance)

Sampling: 279511 pooled, 499 assessed, 140 judged relevant, 349 non-relevant, 10 gray, “C”=1.62, **Est. Rel.**: 43945.8

Final Boolean Result Size (B): 103, **Est. Recall**: 0.1%, **Est. Precision**: 38.3%

Participant High Recall@B: 0.2% (IowaSL0706), **Median Recall@B**: 0.1%

Participant High Recall@25000: 39.4% (UMKC1), **Median Recall@25000**: 14.7%

5 Deepest Sampled Relevant Documents: bts05f00-3598.7 (UMKC1-20802), ypn31e00-3594.6 (otL07fbe-20717), wgp97d00-3427.9 (UMKC5-17662), gyd20e00-3425.8 (UIowa07LegE5-17627), wpc45c00-3297.4 (UIowa07LegE3-15687)

Topic 97 (2007-D-9)

Request Text: Submit all documents that relate to any plans of, interest in, or efforts undertaken for any acquisition, divestiture, joint venture, alliance, or merger of any kind within or related to the property and casualty insurance industry.

Initial Proposal by Defendant: (plan OR interest OR effort) AND (acquisition OR divestiture OR "joint venture" OR alliance OR merger) AND ("property insurance" OR "casualty insurance")

Rejoinder by Plaintiff: (acquisition OR divestiture OR venture OR alliance OR merger) AND insurance

Final Negotiated Boolean Query: (acquisition OR divestiture OR "joint venture" OR alliance OR merger) AND ((property OR casualty) AND insurance)

Sampling: 256752 pooled, 499 assessed, 90 judged relevant, 404 non-relevant, 5 gray, “C”=2.36, **Est. Rel.**: 9032.0

Final Boolean Result Size (B): 13296, **Est. Recall**: 29.4%, **Est. Precision**: 18.1%

Participant High Recall@B: 33.0% (CMUL07ibs), **Median Recall@B**: 10.1%

Participant High Recall@25000: 71.7% (otL07pb), **Median Recall@25000**: 11.6%

5 Deepest Sampled Relevant Documents: bib83c00-2696.3 (otL07pb-13811), cht55f00-1758.8 (CMUL07ibs-12259), rnr35f00-1451.4 (ursinus4-7542), czr93f00-1100.7 (otL07pb-4432), oht84f00-779.0 (UIowa07LegE3-2600)

Topic 98 (2007-D-10)

Request Text: Submit all documents that describe the policies and procedures relating to the retention and destruction of documents (hard copy or electronic) for any company in the property and casualty insurance industry.

Initial Proposal by Defendant: record w/2 (schedule OR retention OR destruction) AND ("property insurance" OR "casualty insurance")

Rejoinder by Plaintiff: (schedule OR retention OR destr!) AND insurance

Final Negotiated Boolean Query: (record w/5 (schedule OR retention OR destr!)) AND ((property OR casualty) AND insurance)

Sampling: 256036 pooled, 499 assessed, 100 judged relevant, 385 non-relevant, 14 gray, “C”=1.33, **Est. Rel.**: 26640.9

Final Boolean Result Size (B): 682, **Est. Recall**: 0.5%, **Est. Precision**: 19.2%

Participant High Recall@B: 2.3% (wat4feed), **Median Recall@B**: 0.4%

Participant High Recall@25000: 39.1% (fdwim7sl), **Median Recall@25000**: 15.4%

5 Deepest Sampled Relevant Documents: lbh20f00-3725.4 (otL07pb-19437), yav99c00-3613.9 (catchup0701p-17338), pvl48d00-3389.9 (ursinus3-14001), aql40f00-2878.1 (UIowa07LegE5-9020), qfb94a00-2747.0 (UMass12-8108)

Topic 99 (2007-D-11)

Request Text: Submit all documents describing natural disasters leading to claims handled by the property and casualty insurance industry.

Initial Proposal by Defendant: "natural disaster" AND ("property insurance" OR "casualty insurance")

Rejoinder by Plaintiff: ("natural disaster!" OR devastation OR catastroph!) AND insurance

Final Negotiated Boolean Query: ("natural disaster!" OR earthquake! OR fire! OR flood!) AND ((property OR casualty) AND insurance)

Sampling: 286700 pooled, 498 assessed, 84 judged relevant, 414 non-relevant, 0 gray, “C”=2.36, **Est. Rel.**: 7484.0

Final Boolean Result Size (B): 19716, **Est. Recall**: 20.6%, **Est. Precision**: 8.2%

Participant High Recall@B: 52.1% (CMUL07o3), **Median Recall@B**: 31.1%

Participant High Recall@25000: 55.0% (CMUL07o1), **Median Recall@25000**: 33.8%

5 Deepest Sampled Relevant Documents: pzl46e00-3320.6 (UIowa07LegE3-23332), fex21c00-1283.7 (fdwim7sl-4492), frz84a00-695.5 (otL07fbe-1993), oqc77e00-613.0 (UMKC5-1713), yhc77e00-440.1 (UMKC5-1169)

Topic 100 (2007-D-12)

Request Text: Submit all documents representing or referencing a formal statement by a CEO of a tobacco company describing a company merger or acquisition policy or practice.

Initial Proposal by Defendant: "formal statement" AND (CEO OR C.E.O. OR "Chief Executive Officer") AND (merger OR acquisition)

Rejoinder by Plaintiff: (CEO OR C.E.O. OR chief OR head) AND (merger OR acquisition)

Final Negotiated Boolean Query: (CEO OR C.E.O. OR "Chief Executive Officer") AND (merger OR acquisition)

Sampling: 310268 pooled, 497 assessed, 93 judged relevant, 404 non-relevant, 0 gray, “C”=1.31, **Est. Rel.**: 8710.0

Final Boolean Result Size (B): 11480, **Est. Recall**: 43.9%, **Est. Precision**: 31.6%

Participant High Recall@B: 45.0% (wat5nofeed), **Median Recall@B**: 19.7%

Participant High Recall@25000: 61.2% (wat1fuse), **Median Recall@25000**: 24.0%

5 Deepest Sampled Relevant Documents: qjh54d00-3378.7 (UIowa07LegE1-13650), bek21f00-719.2 (otL07pb-1372), hzl04e00-651.3 (ursinus4-1191), oqa08c00-528.8 (IowaSL0707-900), rrk60d00-518.3 (wat7bool-877)

Topic 101 (2007-D-13)

Request Text: Submit all documents specifically referencing a lawsuit filed against a named property and casualty insurance company (as either sole or joint defendant).

Initial Proposal by Defendant: (lawsuit OR "complaint filed") AND ("property insurance" OR "casualty insurance")

Rejoinder by Plaintiff: (lawsuit OR complaint OR pleading) AND insurance

Final Negotiated Boolean Query: (lawsuit OR "complaint filed") AND ((property OR casualty)) AND insurance

Sampling: 204551 pooled, 1000 assessed, 184 judged relevant, 785 non-relevant, 31 gray, “C”=6.68, **Est. Rel.**: 8950.9

Final Boolean Result Size (B): 6008, **Est. Recall**: 22.0%, **Est. Precision**: 27.4%

Participant High Recall@B: 30.8% (wat8gram), **Median Recall@B**: 10.0%

Participant High Recall@25000: 81.5% (wat3desc), **Median Recall@25000**: 25.3%

5 Deepest Sampled Relevant Documents: vhv39e00-1403.1 (wat4feed-13029), ihj31e00-492.2 (IowaSL0707-5569), rvl21f00-484.4 (wat3desc-5421), kon90d00-480.0 (wat8gram-5340), dsq44a00-478.4 (UMKC5-5310)

The 10 assessed Relevance Feedback topics are listed next. The summary information differs from that given earlier for the Ad Hoc topics as follows:

Topic : The 10 topics were selected from 2006, whose numbers ranged from 6 to 51. There were 5 complaints in 2006 (labelled A, B, C, D and E).

Initial Proposal by Defendant and Final Negotiated Boolean Query :

The “Rejoinder by Plaintiff” is not listed because it was usually the same as the Final Negotiated Boolean Query in 2006.

Sampling and Est. Rel. : The number of pooled documents includes not just the the residual output of the relevance feedback runs but also all of the documents submitted by the Interactive runs and the special oldrel07 and oldnon07 runs. The number presented to the assessor, the number the assessor judged relevant, the number the assessor judged non-relevant, and the number the assessor left as "gray" are based on the full pool and hence includes rejudging of some documents judged in 2006 (particularly from the oldrel07 and oldnon07 runs). But the "Est. Rel." is just the estimated number of *residual* relevant documents in the pool for the topic (i.e., the number of relevant documents after those judged in 2006 are removed). Hence, unlike for the Ad Hoc topics, "Est. Rel." can be (and sometimes is) lower than "judged relevant".

Final Boolean Result Size B$_r$, Est. Recall and Est. Precision : "B$_r$" is the number of documents matching the final negotiated Boolean query after the documents judged in 2006 are removed; for 2 topics, B$_r$ exceeded 25,000. "Est. Recall" and "Est. Precision" are for the residual Boolean result set.

Feedback High Recall@B$_r$ and Median Recall@B$_r$: Only the 5 runs which used feedback (i.e., the runs which made use of the 2006 judgments) are considered for this listing.

Feedback High Recall@25000 and Median Recall@25000 : Again, only the 5 runs which used feedback are considered for this listing.

5 Deepest Sampled Relevant Documents : Only residual relevant documents are listed. The ranks following the run identifiers are residual ranks. For Interactive runs, all documents were assigned rank 1.

Topic 7 (2006-A-2)

Request Text: All documents discussing, referencing, or relating to company guidelines, strategies, or internal approval for placement of tobacco products in movies that are mentioned as G-rated.

Initial Proposal by Defendant: `(guidelines OR strategies OR "internal approval") AND placement AND "G-rated movie"`

Final Negotiated Boolean Query: `((guide! OR strateg! OR approv!) AND (place! or promot!)) AND (("G-rated" OR "G rated" OR family) W/5 (movie! OR film! OR picture!))`

Sampling: 119645 pooled, 500 assessed, 170 judged relevant, 318 non-relevant, 12 gray, "C"=2.58, **Est. Rel.**: 1280.2

Final Boolean Result Size (B$_r$): 425, **Est. Recall**: 0.9%, **Est. Precision**: 4.8%

Feedback High Recall@B$_r$: 16.7% (sab07legrf1), **Median Recall@B$_r$**: 10.3%

Feedback High Recall@25000: 95.0% (CMU07RFBSVME), **Median Recall@25000**: 88.4%

5 Deepest Sampled Relevant Documents: mak24d00-640.7 (CMU07RSVMNP-1896), lpd41a00-194.5 (CMU07RSVMNP-522), gvc91c00-55.7 (CMU07RFBSVME-416), czc42e00-50.0 (sab07legrf1-313), vwj35c00-44.2 (CMU07RFBSVME-238)

Topic 8 (2006-A-3)

Request Text: All documents discussing, referencing or relating to company guidelines, strategies, or internal approval for placement of tobacco products in live theater productions.

Initial Proposal by Defendant: `(guidelines OR strategies OR "internal approval") AND placement AND ("live theater" OR "live theatre")`

Final Negotiated Boolean Query: `((guide! OR strateg! OR approv!) AND (place! or promot!) AND (live W/5 (theatre OR theater OR audience"))`

Sampling: 119011 pooled, 244 assessed, 100 judged relevant, 143 non-relevant, 1 gray, "C"=2.12, **Est. Rel.**: 10983.9

Final Boolean Result Size (B$_r$): 623, **Est. Recall**: 1.7%, **Est. Precision**: 37.1%

Feedback High Recall@B$_r$: 4.2% (sab07legrf3), **Median Recall@B$_r$**: 2.5%

Feedback High Recall@25000: 50.0% (sab07legrf3), **Median Recall@25000**: 17.4%

5 Deepest Sampled Relevant Documents: cyy57d00-1942.2 (sab07legrf3-6733), jwf04e00-1695.8 (sab07legrf2-5440), eif71c00-1166.4 (CMU07RSVMNP-3225), ahw04c00-1129.4 (sab07legrf3-3093), cqs81e00-1015.9 (CMU07RBase-2703)

Topic 13 (2006-A-9)

Request Text: All documents to or from employees of a tobacco company or tobacco organization referring to the marketing, placement, or sale of chocolate candies in the form of cigarettes.

Initial Proposal by Defendant: `(marketing OR placement OR sale) AND "chocolate cigarettes" AND candy`

Final Negotiated Boolean Query: `(cand! OR chocolate) w/10 cigarette!`

Sampling: 123630 pooled, 250 assessed, 28 judged relevant, 212 non-relevant, 10 gray, "C"=3.01, **Est. Rel.**: 288.8

Final Boolean Result Size (B$_r$): 20240, **Est. Recall**: 99.3%, **Est. Precision**: 1.4%

Feedback High Recall@B$_r$: 100.0% (CMU07RFBSVME and 1 other), **Median Recall@B$_r$**: 76.7%

Feedback High Recall@25000: 100.0% (CMU07RFBSVME and 1 other), **Median Recall@25000**: 76.7%

5 Deepest Sampled Relevant Documents: egg72c00-153.4 (CMU07RFBSVME-480), exg09c00-67.3 (CMU07RFBSVME-206), xej32e00-35.2 (otRF07fb-107), akq55a00-12.3 (sab07legrf3-37), oyz23a00-4.3 (CMU07RFBSVME-13)

Topic 26 (2006-C-2)

Request Text: All documents discussing or referencing retail prices of tobacco products in the city of San Diego.

Initial Proposal by Defendant: `"retail prices" AND tobacco AND California`

Final Negotiated Boolean Query: `((retail OR net) w/2 pric!) AND ("San Diego" or ("S.D." w/3 Calif!))`

Sampling: 104952 pooled, 250 assessed, 95 judged relevant, 152 non-relevant, 3 gray, "C"=2.11, **Est. Rel.**: 15466.3

Final Boolean Result Size ($\mathbf{B}_r$): 2301, **Est. Recall**: 3.1%, **Est. Precision**: 20.9%

Feedback High Recall@$\mathbf{B}_r$: 11.2% (sab07legrf3), **Median Recall@$\mathbf{B}_r$**: 3.9%

Feedback High Recall@25000: 71.0% (sab07legrf3), **Median Recall@25000**: 47.6%

5 Deepest Sampled Relevant Documents: dkm65e00-2785.0 (sab07legrf2-13265), mqp25d00-2650.1 (sab07legrf3-11898), ubo68c00-1820.0 (CMU07RFBSVME-6038), gcd65e00-1670.5 (CMU07RBase-5293), lpk24d00-1055.2 (sab07legrf2-2822)

Topic 27 (2006-C-3)

Request Text: All documents discussing or relating to the placement of product logos at events held in the State of California.

Initial Proposal by Defendant: `"product placement" AND "logos" AND California`

Final Negotiated Boolean Query: `("product placement" OR advertis! OR market! OR promot!) AND (logo! OR symbol OR mascot OR marque OR mark) AND (California OR cal. OR calif. OR "CA")`

Sampling: 335971 pooled, 249 assessed, 120 judged relevant, 129 non-relevant, 0 gray, "C"=1.96, **Est. Rel.**: 23229.7

Final Boolean Result Size ($\mathbf{B}_r$): 127525, **Est. Recall**: 36.7%, **Est. Precision**: 6.9%

Feedback High Recall@$\mathbf{B}_r$: 64.0% (sab07legrf2), **Median Recall@$\mathbf{B}_r$**: 49.1%

Feedback High Recall@25000: 59.1% (CMU07RSVMNP), **Median Recall@25000**: 20.0%

5 Deepest Sampled Relevant Documents: ofg48e00-3543.2 (CMU07RSVMNP-23835), uzn25d00-3455.1 (CMU07RBase-21917), ber19c00-3292.7 (sab07legrf2-18900), urt66d00-3156.6 (CMU07RSVMNP-16781), dah15d00-2441.8 (CMU07RSVMNP-9354)

Topic 30 (2006-C-6)

Request Text: All documents discussing or referencing the California Cartwright Act.

Initial Proposal by Defendant: `"California Cartwright Act"`

Final Negotiated Boolean Query: `California w/3 (antitrust OR monopol! OR anticompetitive OR restraint OR "unfair competition" OR "Cartwright")`

Sampling: 125617 pooled, 250 assessed, 24 judged relevant, 226 non-relevant, 0 gray, "C"=2.34, **Est. Rel.**: 7.0

Final Boolean Result Size ($\mathbf{B}_r$): 202, **Est. Recall**: 28.6%, **Est. Precision**: 1.8%

Feedback High Recall@$\mathbf{B}_r$: 57.1% (CMU07RFBSVME and 1 other), **Median Recall@$\mathbf{B}_r$**: 42.9%

Feedback High Recall@25000: 100.0% (CMU07RFBSVME and 2 others), **Median Recall@25000**: 100.0%

5 Deepest Sampled Relevant Documents: idc90e00-1.0 (sab07legrf1-5), zce78c00-1.0 (CMU07RSVMNP-5), phh71c00-1.0 (CMU07RSVMNP-3), dzk44c00-1.0 (CMU07RBase-3), pbx64d00-1.0 (CMU07RBase-1)

Topic 34 (2006-D-1)

Request Text: All documents discussing or referencing payments to foreign government officials, including but not limited to expressly mentioning "bribery" and/or "payoffs."

Initial Proposal by Defendant: `(bribery OR payoffs) AND payments AND "foreign government officials"`

Final Negotiated Boolean Query: `(payment! OR transfer! OR wire! OR fund! OR kickback! OR payola OR grease OR bribery OR payoff!) AND (foreign w/5 (official! OR ministr! OR delegat! OR representative!))`

Sampling: 122598 pooled, 248 assessed, 105 judged relevant, 140 non-relevant, 3 gray, "C"=2.41, **Est. Rel.**: 20113.0

Final Boolean Result Size ($\mathbf{B}_r$): 2380, **Est. Recall**: 1.8%, **Est. Precision**: 16.5%

Feedback High Recall@$\mathbf{B}_r$: 6.5% (CMU07RSVMNP), **Median Recall@$\mathbf{B}_r$**: 2.0%

Feedback High Recall@25000: 54.7% (CMU07RSVMNP), **Median Recall@25000**: 16.3%

5 Deepest Sampled Relevant Documents: rut15e00-2950.9 (CMU07RSVMNP-17353), yph30a00-2910.5 (CMU07RBase-16784), oyf37c00-2719.0 (sab07legrf1-14364), uva40f00-2686.7 (sab07legrf3-13995), nnj14e00-2587.3 (CMU07RSVMNP-12922)

Topic 37 (2006-D-4)

Request Text: All documents relating to defendants' tobacco advertising, marketing or promotion plans in China's capital.

Initial Proposal by Defendant: `(advertising OR marketing OR "promotion plans") AND (China OR Beijing)`

Final Negotiated Boolean Query: `(advertis! OR market! OR promot! OR encourag! OR incentiv!) AND (China OR Beijing OR Peking)`

Sampling: 149493 pooled, 250 assessed, 74 judged relevant, 175 non-relevant, 1 gray, "C"=2.85, **Est. Rel.**: 7086.3

Final Boolean Result Size ($\mathbf{B}_r$): 38723, **Est. Recall**: 59.0%, **Est. Precision**: 13.6%

Feedback High Recall@$\mathbf{B}_r$: 50.6% (sab07legrf1), **Median Recall@$\mathbf{B}_r$**: 49.2%

Feedback High Recall@25000: 50.6% (sab07legrf1), **Median Recall@25000**: 49.2%

5 Deepest Sampled Relevant Documents: rgd60a00-2384.7 (CMU07RSVMNP-12994), aer19e00-1178.8 (sab07legrf2-4396), jmy90d00-1100.0 (otRF07fb-4019), awk95c00-1018.2 (CMU07RBase-3644), ziv19e00-519.1 (sab07legrf1-1651)

Topic 45 (2006-E-4)

Request Text: All documents that refer or relate to pigeon deaths during the course of animal studies.

Initial Proposal by Defendant: `"animal studies" AND "pigeon deaths"`

Final Negotiated Boolean Query: `(research OR stud! OR "in vivo") AND pigeon AND (death! OR dead OR die! OR dying)`

Sampling: 112239 pooled, 498 assessed, 91 judged relevant, 400 non-relevant, 7 gray, "C"=4.97, **Est. Rel.**: 83.2

Final Boolean Result Size ($\mathbf{B}_r$): 2507, **Est. Recall**: 70.0%, **Est. Precision**: 2.4%

Feedback High Recall@$\mathbf{B}_r$: 97.6% (sab07legrf2), **Median Recall@$\mathbf{B}_r$**: 94.5%

Feedback High Recall@25000: 100.0% (CMU07RSVMNP and 2 others), **Median Recall@25000**: 100.0%

5 Deepest Sampled Relevant Documents: qwl16d00-16.0 (CMU07RSVMNP-82), jbr10a00-15.4 (otRF07fb-79), ivw00a00-8.3 (CMU07RSVMNP-42), wzn20a00-7.9 (otRF07fb-40), phg81d00-3.6 (otRF07fb-18)

Topic 51 (2006-E-10)

Request Text: All documents referencing or regarding lawsuits involving claims related to memory loss.

Initial Proposal by Defendant: `(lawsuits OR "tort claims") AND "memory loss"`

Final Negotiated Boolean Query: `((memory w/2 loss) OR amnesia OR Alzheimer! OR dementia) AND (lawsuit! OR litig! OR case OR (tort w/2 claim!) OR complaint OR allegation!)`

Sampling: 108434 pooled, 499 assessed, 83 judged relevant, 400 non-relevant, 16 gray, "C"=4.01, **Est. Rel.**: 65.7

Final Boolean Result Size ($\mathbf{B}_r$): 6927, **Est. Recall**: 84.8%, **Est. Precision**: 0.8%

Feedback High Recall@$\mathbf{B}_r$: 84.8% (CMU07RFBSVME), **Median Recall@$\mathbf{B}_r$**: 23.9%

Feedback High Recall@25000: 84.8% (CMU07RFBSVME), **Median Recall@25000**: 46.7%

5 Deepest Sampled Relevant Documents: qcm10d00-7.7 (CMU07RBase-31), war78e00-1.0 (randomRF07-4), ptn85d00-1.0 (uw07T2-1), bee61e00-1.0 (uw07T2-1), sfw87e00-1.0 (liu07-1)

---

5 Workshop Discussion

There were several opportunities for interaction among the participants from the research and legal communities during the conference, culminating in the Thursday, November 8 workshop which discussed future plans for the track (which will continue for a 3rd year in 2008).

At the workshop, two smaller document sets were considered for 2008. One option was a collection of State Department cables from the 1970's, which would be a cleaner collection to work with (e.g., no OCR issues), but there were concerns about whether the legal community would accept results based on it. The other option was an Enron collection, which would feature email resembling modern e-discovery scenarios, but it was considered a difficult collection to work with (e.g., attachments in proprietary formats). It was decided to continue in 2008 with the same IIT CDIP collection as the past couple years, particularly since there were a lot of new participants in 2007 who would like the chance to fully focus on research issues in 2008 rather than deal with the details of using a new collection.

There were concerns raised at the workshop about the appropriateness of the Recall@B measure which was used as the primary measure in 2007; e.g., the real goal of discovery is to produce the set of relevant documents, not just to maximize success at a particular given size B. (We followup on the choice of measure in the next section.)

More focus on relevance feedback was suggested at the workshop, both to encourage more use of metadata (e.g., author, Bates number) and to enrich the relevance judgments for past topics to further improve their re-usability. Deeper and denser assessing was also suggested, even if it meant fewer new topics.

A proposal also discussed among track coordinators before and during the workshop concerned whether in future years the Legal Track should introduce and evaluate the concept of "highly relevant" documents, as a third category for purposes of assessment along with not relevant and relevant. The problem of isolating a true set of "hot" or "material" documents for use in later phases of discovery (e.g., depositions) and at trial, amongst a large universe of merely potentially tangentially relevant documents, remains a key concern for the legal profession. This issue will be explored further in Year 3 of the track.

We look forward to continuing the discussion in 2008!

5.1 Post-Workshop Analysis

After the conference, we analyzed the Ad Hoc runs from the perspective of trying to produce a set as close as possible to the desired set of R relevant documents. In particular, we looked at the F1 measure which combines recall and precision into one measure (F1 = 2*Prec*Recall / (Prec+Recall)).

The reference Boolean run averaged an F1 of 0.14 over the 43 topics, whereas if we cutoff the Ad Hoc runs at their top-ranked R retrieved (where R is the estimated number of relevant documents, rounding up fractional estimated R values to the next integer) several Ad Hoc runs scored higher (to a high of 0.22 (otL07frw); the median was also 0.14). Hence, if the Ad Hoc runs can pick a good cutoff value, they apparently can produce a closer set to the optimal set of R documents than the reference Boolean run's set of B documents, taking both recall and precision into account.

This result suggests that we should enhance the Ad Hoc task in 2008 to require each system to additionally specify a cutoff value K for each topic. (Unfortunately, in 2007, we did not ask the Ad Hoc systems to specify a cutoff value before R was known.) A measure which balances recall and precision (such as F1) could then be used to evaluate whether automated approaches can produce a set of K documents closer to the optimal set of R relevant documents than the reference Boolean query result set (for which K=B). We would still ask the systems to submit their top-25,000 ranked documents (or whatever the agreed limit is in 2008) to enrich the pools and enable post hoc analysis of different choices of K.

6 Conclusion

In its second year, the TREC Legal Track made several advances. The Ad Hoc task developed a much deeper sampling approach to more accurately estimate recall and precision and evaluated a wider variety of automated search techniques thanks to a doubling in participation. A separate Interactive task was created for studying the effectiveness of "expert" searchers. A new Relevance Feedback task was created to study automated ways of making use of judgments from an initial sample. Baseline results for each task were established and several resources are now available to support further study going forward.

For the Ad Hoc task, 50 new topic statements, i.e., requests for documents with associated negotiated Boolean queries, were created. 12 research teams used a wide variety of (mostly automated) techniques to search the IIT CDIP collection (a complex collection of almost 7 million scanned documents) and submit a total of 68 result sets of 25,000 top-ranked documents for each topic. These submissions were pooled, producing a set of approximately 300,000 documents per topic. A new sampling scheme was used to select between 500 and 1000 documents from the pool for each topic. Volunteers from the legal community assessed 43 of the 50 result samples in time for reporting the results at the conference. Based on the samples, we can estimate that there were on average almost 17,000 relevant documents per topic, and that this number varied considerably by topic (from a low of 18 to a high of more than 77,000).

The deep sampling allows us to estimate the recall and precision of the final negotiated Boolean query more accurately than before. On average (over the 43 topics), the reference Boolean query found just 22% of the relevant documents that are estimated to exist. Its precision averaged 29%. Again, these numbers varied considerably by topic (the recall ranged from 0% to 100%, while the precision ranged from 0% to 97%). It is quite striking that, on average per topic, 78% of the relevant documents were only found by participant research techniques and not by the reference Boolean query.

Surprisingly, when recall was estimated at depth B, where B is the number of documents matched by the reference Boolean query, no system participating in the Ad Hoc task submitted results that improved over the reference Boolean run (on average), despite the systems' collective success at finding relevant documents missed by the Boolean query. However, post hoc analysis using the F1 measure (which balances recall and precision) found that the Ad Hoc systems potentially can produce a set of results closer to the optimal set of R relevant documents than the reference Boolean result set. Unfortunately, this latter possibility was not properly evaluated in 2007 because the systems were not asked to specify a cutoff value before R was known (where R is the estimated number of relevant documents). We should consider refining the methodology in 2008 to require each system to specify a cutoff value K for each topic for targeting a measure (such as F1) which balances the demand for recall with the cost of reviewing unresponsive documents.

A new Interactive task was created in 2007 to followup on an interesting result from 2006, which was that the sole expert searcher achieved a higher mean R-precision than any of the automated runs in 2006 (albeit based on the shallower sampling used in 2006, which underestimated R considerably). In 2007, 8

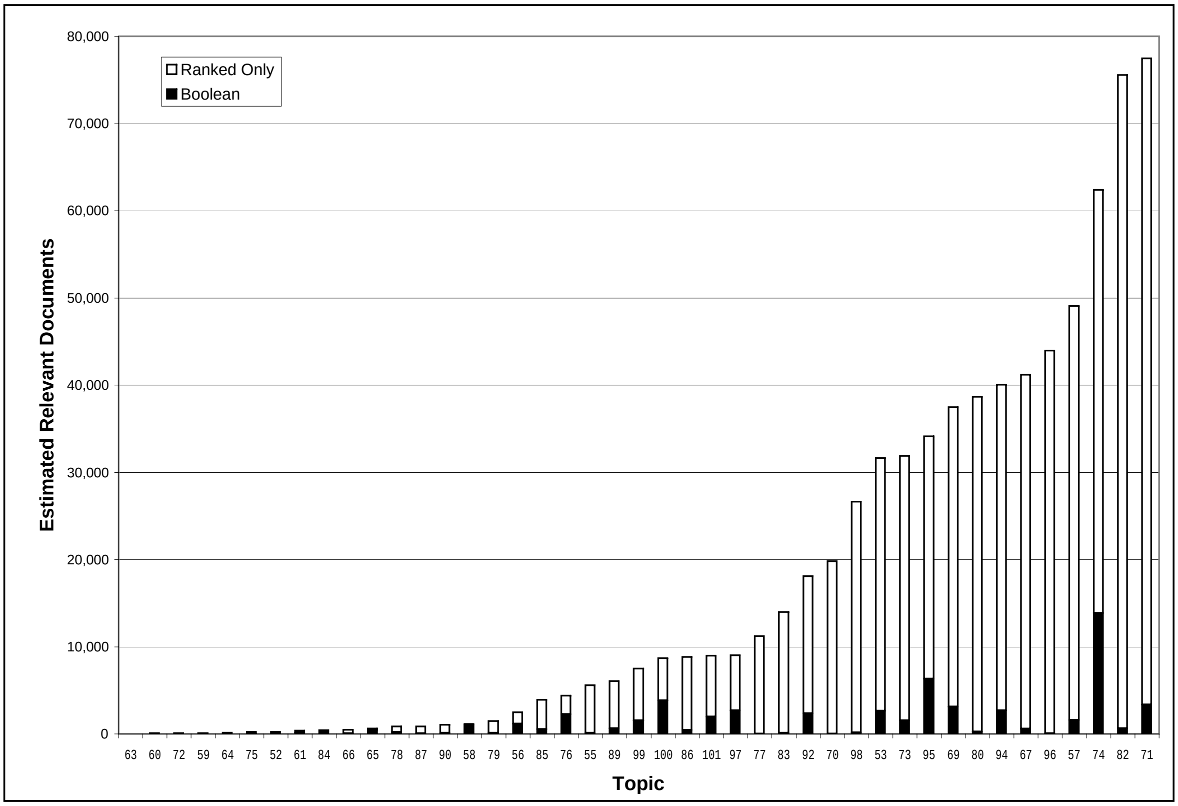

Figure 1: Estimated relevant documents found by the reference Boolean query (black) and found only by one or more ranked systems (white) for the 43 Ad Hoc topics.

teams from 3 sites took up the Interactive challenge. Some teams invested several hours per search topic, and most teams completed just 3 topics. It was found that there was substantial variation in the results of the participating teams, but most of them outscored the reference Boolean query in the task's utility measure for each of the 3 topics. This result is another encouraging one for expert searching, albeit one with many caveats. For instance, the participating teams in 2007 were limited to submitting 100 documents per topic (as was the sole expert searcher in 2006). We intend to remove this limit in 2008 so that the experts' ability to recall much larger numbers of relevant documents can be evaluated.

In the new Relevance Feedback task of 2007, 10 of the previous year's Ad Hoc topics were re-used. Participants were encouraged to use the previous year's document assessments as feedback to improve their results. 3 teams submitted a total of 8 runs, including 5 feedback runs. Residual evaluation was used, i.e., documents judged in the previous year were removed from the result sets before evaluating. The deep sampling approach of the Ad Hoc task was likewise applied to the Relevance Feedback task. An encouraging result from this pilot study was that the median of the 5 feedback runs found more relevant documents than the reference Boolean run by depth B_r for 7 of the 10 topics (where B_r is the number of documents matched by the reference Boolean query after removing documents judged the previous year), in contrast with the Ad Hoc task in which the median run outscored the reference Boolean run at depth B for just 8 of the 43 topics. We hope to run a larger study of Relevance Feedback (more test topics and more participants) in 2008.

The evaluation of e-discovery approaches remains a daunting challenge. The findings so far are very preliminary. Automated approaches can improve upon the recall of negotiated Boolean results, but typically at the expense of reviewing additional documents. Experts can improve upon automated approaches, but they also vary a lot in performance. Feedback approaches are promising, but a larger study is needed. We are heartened that so many volunteers have contributed to the track's endeavours in 2006 and 2007 and look forward to working with everyone to make further advances in 2008.

Acknowledgments

This track would not have been possible without the efforts of a great many people. Our heartfelt thanks go to Ian Soboroff for creating the relevance assessment system; to the dedicated group of pro bono relevance assessors and the pro bono coordinators at the participating law schools; to Conor Crowley (DOAR Consulting), Joe Looby (FTI Consulting), Stephanie Mendelsohn (Genentech), and the team from H5 (Todd Elmer, Bruce Hedin, Jim Donahue and Michelle Luke) for their invaluable assistance with complaint drafting, topic formulation, and participating in Boolean negotiations; and to Richard Braman, Executive Director of The Sedona Conference®, for his continued support of the TREC Legal Track.

References

[1] TREC Legal Track Home Page. http://trec-legal.umiacs.umd.edu/.

[2] James Allan, Ben Carterette, Javed A. Aslam, Virgil Pavlu, Blagovest Dachev, and Evangelos Kanoulas. Million query track 2007 overview. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*, November 2007. http://trec.nist.gov.

[3] Jason R. Baron, David D. Lewis, and Douglas W. Oard. TREC-2006 legal track overview. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, November 2006. http://trec.nist.gov.

[4] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. Bias and the limits of pooling. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 619–620, 2006.

[5] Stefan Büttcher, Charles L. A. Clarke, and Ian Soboroff. The TREC 2006 terabyte track. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, November 2006. http://trec.nist.gov.

[6] The Sedona Conference. The Sedona conference best practices commentary on the use of search and information retrieval methods in e-discovery. *The Sedona Conference Journal*, pages 189–223, 2007.

[7] Disability Rights Council of Greater Wash. v. Wash. Metro. Area Transit Auth. (D.D.C.). 2007 WL 1585452.

[8] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. Building a test collection for complex document information processing. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 665–666, 2006.

[9] George L. Paul and Jason R. Baron. Information inflation: Can the legal system adapt? *Richmond Journal of Law and Technology*, 13(3), 2007.

[10] H. Schmidt, K. Butter, and C. Rider. Building digital tobacco document libraries at the university of california, san francisco library/center for knowledge management. *D-Lib Magazine*, 8(2), 2002.

[11] Stephen Tomlinson. Early Precision Measures: Implications from the Downside of Blind Feedback. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 705–706, 2006.

[12] Stephen Tomlinson. Experiments with the negotiated boolean queries of the TREC 2006 legal discovery track. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, November 2006. http://trec.nist.gov.

[13] Williams v. Taser Intern, Inc. (N.D. Ga.). 2007 WL 1630875.

[14] Emine Yilmaz and Javed A. Aslam. Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proceedings of the 15th International Conference on Information and Knowledge Management (CIKM)*, pages 102–111, 2006.

[15] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, 1998.