Recommind at TREC 2011 Legal Track

Peter Zeinoun[1], Aaron Laliberte[1], Jan Puzicha[1], Howard Sklar[1], Craig Carpenter[1]

[1]Recommind

INTRODUCTION

The TREC 2011 Legal Track Learning Task (the "2011 Task") evaluated each of approximately 685,000 documents from the Enron data set for responsiveness to one or more requests for production. The 2011 Task included three new requests for production (the "Topics" and each a "Topic")). In this paper, we describe the approaches used to conduct the review by the Recommind team and report the scores for each of the three Topics.

2011 TASK

Overview

To conduct the review, Axcelerate® Review and Analysis v4.2 software (the "Axcelerate System") was installed on a virtual machine to which team members were provided access. The Enron data set was loaded in a .txt format rather than native language to make the review process more efficient and to accommodate reporting requirements. Next, pre-processing and learning creation were performed, and the results were used to set up relevancy scores and extraction.

Initially, the team began identifying a "seed set" of documents for each Topic---a set of documents that would later be used in the Axcelerate System's Predictive Coding process---through various search and entity extraction methodologies including keywords, phrase extraction, and concept searches. Relevant documents were mined for additional terms that could be used to enhance the efficacy of the search. The team then used additional analytics within the Axcelerate System to examine different documents that contained responsive keywords for each Topic and at times all Topics, applying training and relevancy analysis to identify various document sets in different ways. For example, the Axcelerate System's Predictive Analytics automatically categorized documents into computer-generated "buckets" based not just on keyword frequency but on conceptual meaning as well, irrespective of individual keywords.
The three Topics identified by TREC provided an opportunity to teach reviewers the key concepts and examples that applied to all Topics. This enabled reviewers to consider each document's relevancy in parallel, rather than viewing each Topic in multiple, consecutive (and therefore time-consuming) review processes.

Predictive Coding was employed to identify documents, filter them, run training on the results and identify more documents for review. When the filter structure and system were set up, the team moved forward with the actual 2011 Task.

TREC Topic Authorities

TREC designated senior litigators as Topic Authorities (the "TAs") to interpret the requests for production (i.e. Topic) and to determine the responsiveness of documents according to that interpretation. Timeliness of responses was determined by the availability and capacity of each TA. The goal was to have TAs provide responses within 48 hours, for up to 100 documents, per topic, per team. As will be seen, such expected TA responsiveness was unfortunately not seen with respect to Topic 403.

Topic 401

Request 401 sought, "All documents or communications that describe, discuss, refer to, report on, or relate to the design, development, operation, or marketing of EnronOnline, or any other online service offered, provided, or used by the Company (or any of its subsidiaries, predecessors, or successors-in-interest), for the purchase, sale, trading, or exchange of financial or other instruments or products, including but not limited to, derivative instruments, commodities, futures, and swaps."

Guidance

The identified TA indicated that Topic 401 is limited to financial transactions but, as the email below from the TA explained, includes "commodities." The team asked whether trading in energy, as long as the trade occurs online, would be responsive. The TA's response to this inquiry follows:

> *"During the Kick-Off call this week someone asked if the definition of "commodity" (in the phrase "financial instruments/financial products including but not limited to derivative instruments/commodities/futures/swaps") included energy and in particular electricity, oil and gas. EnronOnline was an online trading exchange where third parties bought and sold various financial instruments which included gas, oil and electricity units at quoted prices. EnronOnline was not used for consumer transactions.*
>
> *Therefore, any information related to the origin, design, operation or marketing of EnronOnline or any other system used for trading derivatives or similar financial products (which could include gas, oil and electricity) should be considered Responsive, even if that information does not specifically mention trading."*

Recommind's Interpretation

The first restriction applied to Topic 401 was the relationship of 'online' services offered, provided, or used by the company or any of its subsidiaries to sell financial instruments or products. Various forms of expressing 'online', such as the abbreviation EOL (EnronOnline), or synonyms like 'web site', 'electronic', or 'virtual' were employed.

Initial keyword and phrase extraction searches identified additional relevant terms. Second-level relevancy determinations were then used to identify new documents and more search terms. For example, when an Enron employee responsible for the development of EnronOnline was discovered, that name was added to our "list" of search terms.

Our initial interpretation required a more focused definition of 'financial products' and 'commodities'. As an energy company, Enron's documents included a wide spectrum of oil and electricity issues, including using their online network to sell these products. To further complicate this interpretation, there were also online approvals of transactions, but in many cases there was not enough information to determine the nature of the transaction. The primary interpretive challenges for Topic 401 were determining if the underlying transactions were financial instruments, derivatives, or commodities, and if they were executed traditionally or online.

External counsel also played an important role in our interpretation. They not only provided industry knowledge that helped identify Enron subsidiaries, but also identified additional terms of art that could encompass derivatives and other financial instruments.

TA Considerations

The 401 TA recognized that inclusion of energy products like oil and gas as commodities would significantly expand the search. In correspondence during this period, the TREC leadership team also noted that when Topic 401 had been developed, commodities such as oil and gas were not considered relevant; subsequently and at the direction of the Topic 401 TA, however, oil, gas, and electricity were included as being relevant to Topic 401. As a result of this interpretation, the search was expanded to include oil, gas, and electricity.

Topic 402

Request 402 sought, “All documents or communications that describe, discuss, refer to, report on, or relate to whether the purchase, sale, trading, or exchange of over-the-counter derivatives, or any other actual or contemplated financial instruments or products, is, was, would be, or will be legal or illegal, or permitted or prohibited, under any existing or proposed rule(s), regulation(s), law(s), standard(s), or other proscription(s), whether domestic or foreign."

Guidance

The request sought documents relating to ‘financial instruments’, including ‘over-the-counter derivatives’. The TA provided the following guidance on these terms.

> The International Swaps and Derivatives Association defines “derivatives” as follows:
>
> > “A derivative is a risk transfer agreement, the value of which is derived from the value of an underlying asset. The underlying asset could be an interest rate, a physical commodity, a company’s equity shares, an equity index, a currency, or virtually any other tradable instrument upon which parties can agree.”
>
> The International Swaps and Derivatives Association defines “over-the-counter” derivatives as follows:
>
> > “OTC derivatives, which are sometimes called swap agreements or swaps, are negotiated privately between the two parties and then booked directly with each other.”
>
> Note, however, that Topic 402 broadly related to “any other actual or contemplated financial instruments or products.” Thus, OTC derivatives are only one type of financial instrument that would fall within the scope of the Request. However, due to the nature of Enron’s underlying businesses we might expect many of the documents to relate to OTC derivatives specifically, so we deemed that careful attention should be paid when reviewing documents in which an OTC derivative transaction were referenced.
>
> *Note on exchanges*: A reference to a platform, system or exchange used to transact relevant financial instruments could make a document Responsive even if that document does not mention an actual instrument or product being transacted. For example, a document that relates to the legality or regulation of EnronOnline would be Responsive because EnronOnline was used to transact financial instruments.
>
> *Note on consumer transactions*: A cash or consumer credit-card transaction for the purchase of end-user goods or services is Non-Responsive. For example, a document expressing concern about the legality of the actual or contemplated scalping of football tickets is Non-Responsive. Similarly, a document concerning the illegality of the purchase of alcohol by a minor is Non-Responsive.
>
> *Note on gambling and contests*: A wager or gambling transaction that is premised upon a sporting contest or a pure game of chance (such as lottery ticket, blackjack or fantasy football) is Non-Responsive because it is not a “financial instrument” or “financial product,” nor is it considered a derivative under the definition above.
>
> For example, a document expressing concern about the legality of “fantasy football” activities is Non-Responsive. Similarly, an email concerning a raffle or drawing that indicates that the contest is only valid in states where not prohibited by law is Non-Responsive. However, the factor of chance does not alone render a document about a transaction non-responsive. For example, a document concerning the legality of weather derivative products is Responsive because those products are used to hedge or protect various assets against adverse weather conditions and is therefore considered a financial product rather than a gambling contest. Where unsure whether a

transaction was considered a gambling contest or a financial product, we sought to mark the document as Responsive.

Recommind's Interpretation

Like Topic 401, the interpretation of Topic 402 focused on determining if the term 'financial' applied to both 'instruments' and 'products'. The expanded definition, which included 'products', was initially applied. However, upon receiving coding instructions the search was reduced to include only financial instruments.

The issues of internal and contractual compliance also presented an interpretive challenge. For example, the inclusion of internal policies required close examination before it was decided by the Recommind team that Topic 402 was concerned with external, 'public law' standards that were set by legislation, court opinion, or regulation. Similarly, contractual obligations often referenced 'public law' standards or adopted them as a contractual component, but it was determined that Topic 402 was concerned with external public law standards. Therefore, internal policies or contractual obligations were not included.

TA Considerations

The ambiguity over internal and external policies was not addressed by any explicit TA determinations. Unfortunately, an evaluation of TA determinations received as part of the TA submission process seems to indicate some level of inconsistency in TA treatment of this issue as well with the net result being that it could not be definitively determined from TA determinations whether these policy or contractual obligations should appropriately be included or excluded as a relevancy determination.

Topic 403

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

Guidance

The environmental impact needed to come from Enron's activities, not those of other companies. Examples given by the TA consisted of the following:

1. NO_2 emissions from Enron (or subsidiary) power plants were considered relevant
2. A document about an oil spill from an Enron rig, even if there was no discussion about impact, were considered relevant because it came from Enron's activities
3. Discussions about minimum wage were considered not relevant
4. A document which included the full text of an environmental law would be considered not relevant unless it was tied with a company activity

Recommind's Interpretation

A key interpretive challenge focused on the definition of 'environmental impact'. For example, several documents referred to testing performed at various sites, but it was difficult to determine if these tests were related to an environmental impact. Additional interpretive challenges included 1) determining if the mention of a potential contaminant accounts for an environmental impact; 2) determining if CO_2 and NO_2 emissions reports were relevant; and 3) determining if testing for pipeline contaminants referred to purity of product rather than an environmental impact.

The Recommind team's interpretation arrived at conclusions for each of these issues which were never, unfortunately, verified by the TA (as noted below). It was determined that an impact on the environment was required for relevance, and that all Enron subsidiaries would be treated as relevant. This was driven by the TA guidance noting that NO_2 emissions from Enron or subsidiaries were relevant. It was determined that testing for a pipeline's contaminants related to product purity, so these documents were considered not relevant.

TA Considerations

The exact scope of environmental impact necessary to trigger relevance did not seem to be consistent across TA determinations, making it difficult to confirm or deny whether the Recommind interpretation of any impact to the environment was the appropriate standard, or if it had to rise to the level of a violation of environmental law. Similarly, a non-trivial portion of the document collection, which Recommind identified as responsive, did not explicitly identify a subsidiary as Enron, but rather that information was imputed from other documents. Recommind has a particular concern that this imputed knowledge, a strength of human review that can be amplified by machine learning, might not be present in solely machine learning evaluations – especially, as here, where no TA corroboration was provided. Exacerbating this Topic and TA determination inconsistency was a lack of responsiveness of the Topic 403 TA; unclear documents went unresolved as the TA simply did not respond to submissions from the Recommind team, leaving the Recommind team entirely in the dark as to many documents. We certainly understand that TAs volunteer their time for which all those participating in the 2011 Task are grateful. However, these significant issues with respect to consistency and timeliness had an unfortunate materially detrimental effect on the Topic 403 results of the Recommind team, and are not reflective of how Recommind customers utilize the Axcelerate System's Predictive Coding process to great effect on a daily basis.

PATENTED PREDICTIVE CODING WORKFLOW

The patented Predictive Coding process implemented by Recommind was based on three core workflow steps:

> **1. Predictive Analytics**: Predictive Analytics includes the use of keyword, Boolean and concept search and data mining techniques to help a case management team develop understanding of a matter and quickly identify sets (batches) of probative documents for review. These sets are reviewed by the case team and establish seed documents to serve as examples for the Predictive Coding algorithm.
>
> **2. Iterations**: Iterations are multiple occurrences of category training that identify additional documents that are substantively similar to seed documents. Documents identified as being probative of a category during human review using Predictive Analytics are used by the Predictive Coding algorithm as further examples of the documents that belong in that category, enriching the patterns the algorithm iteratively applies to the as-yet uncategorized documents in the corpus. The cycle is as follows:
>
>> The probative seed documents are used as input for a categorization run;
>>
>> The system identifies documents that are substantively similar to the seed set for such category and returns them in ranked order (from "most" like the seed set to "least" like the seed set);
>>
>> The case team reviews/codes the suggested documents, providing further calibration for the Axcelerate System; and
>>
>> All probative seed documents are then 'trained' upon, with the iterations continuing until no more algorithmically similar documents remain.

3. Predictive Sampling: Predictive Sampling is the use of statistical sampling as a quality control process to test and verify the results of Predictive Coding. Statistical sampling provides quantifiable validation that the process used generated sufficiently accurate results. Sampling is used after iterations yield no or a very small number of additional probative documents, meaning very few responsive documents remain uncoded. The process entails selecting a random sample of documents that have not been reviewed and placing them under human evaluation for responsiveness. The results of the review of the sampled documents from the uncoded set can then be extrapolated to the entire uncoded set, thus establishing the number of potentially responsive documents in the entire uncoded set to a degree of statistical certainty.

Quality Control

The review process consisted of three levels of human review, each with a dedicated review team. The levels were:

1. First Level Review – the first level at which documents underwent human review. At this level documents were reviewed for responsiveness and coded as Responsive, Non-Responsive or For Further Review

2. Second Level Review – the second level at which documents underwent human review. Documents contained in the second level review had already undergone first level review. Quality control was conducted at this level by comparing determinations made by the second level reviewer against those made by the first level reviewer. The review process includes random sampling of documents per reviewer and review of false positive and false negative documents

3. Third Level Review – the third and last level at which documents underwent human review. Documents contained in the Third Level Review had already undergone first and second level review. Review at this level was designed to address disagreements between 1) first and second level review, and 2) TA determinations and second level review. Determinations made at this level are considered definitive.

Random Sampling

Random samples of first level review documents were taken per reviewer and reviewed by a second level reviewer. The level of disagreement between first and second level review was then used to measure the accuracy of the first level reviewer. If there was a high percent of discrepancies the root of the discrepancies was investigated and addressed. The frequency of quality control was adjusted according to the findings of the review.

Review of False Positives and Negatives

A review of false positive and false negative documents was conducted as part of the quality control effort.

False positives are defined as documents to which the Predictive Coding algorithm assigned a high probability of responsiveness but were coded as non-responsive during human review. False positives were assigned to second level review and if there was an error on the part of the reviewer the document would be re-coded accordingly. If there was no error on the part of the reviewer the reviewer would analyze the document for unique qualities that could lead to the discovery of similar false positives. A search would then be done based on the identified unique qualities and the results would be batched out for review.

False negatives are defined as documents to which the Predictive Coding algorithm assigned a low probability of responsiveness but were coded as responsive during human review. False positives were assigned to second level review and if there was an error on the part of the reviewer the document would be re-coded accordingly. If there was no error on the part of the reviewer the reviewer would analyze the document for unique qualities that could lead to the discovery of similar false negatives. A search would then be done based on the identified unique qualities and the results would be batched out for review. Newly discovered responsive documents were then added to the seed set.

Sample Size Generation

The error rate of a specific tag (a tag being the identification of a particular document being identified as a member of a particular category of documents responsive to a specific issue for which responsive documents must be produced to satisfy a document request, and the error rate being the rate at which the tag is incorrectly associated with documents from a corpus of documents) is Bernoulli distributed, so it can be easily estimated with a specified degree of confidence within any specified confidence interval by manual review of a set of documents, as long as the sampling method is unbiased. The estimate is simply given by the empirical mean over the sample. The sample size required to be certain about a given error rate can be guaranteed with, for example, 95% probability that the true error rate for a given estimate is bound by:

$$e = \mathrm{p} + \frac{1.645}{2\sqrt{\mathrm{n}}}$$

As long as the batch is sampled in an unbiased fashion, this procedure can be applied to provide an accurate estimate of the error rate for a given set of documents.

As a typical requirement, the error rate on unreviewed documents must be determined with a certain level of confidence within a specific estimation interval. The Axcelerate System provides users with the ability to select the confidence and estimation interval to automatically compute the required sample sizes.

Document Representation

Any categorization technique relies on some document representation, usually in terms of a feature vector describing the document content. Within the Axcelerate System, such features include

- “Bag-of-words”, i.e. counts over word occurrences
- Auto-extracted noun phrases
- Possibly auxiliary meta-data
- Topical features derived from probabilistic latent semantic indexing (PLSI)

Probabilistic Latent Semantic Analysis (PLSA)

PLSA is a patented algorithm that performs a statistical analysis of multi-dimensional word co-occurrences in documents and identifies repeatable contexts, topics or concepts in which a certain group of words occurs. It does not require any manual input in the form of lexicons, thesauri, or topic annotations, but is completely automatic in performing *unsupervised learning*. Through PLSA the Axcelerate System is able to group documents together based upon similar concepts and can do so even in the absence of taxonomies and other categorical information. The outcome of the learning procedure is what we call a statistical *model*, a compressed, quantitative description of the document collection. This conceptual representation can then be used as input for a subsequent categorization step.

The identification of concepts or topics serves two purposes: first, it reveals the potential ambiguity of words by detecting multiple contexts in which they are used. For example, “jaguar” may refer to the animal, the automobile brand, and any number of clubs, products, and businesses; “Java” might refer to the Indonesian island, the programming language, or coffee. Such ambiguities, also called *polysemies*, are automatically identified by PLSA whenever they are present in the source documents.

Second, PLSA learns about *synonyms* and *semantically related words*, i.e., words that are likely to occur in a common context. For example, a document containing the term "car" is likely to contain synonyms like "automobile", "auto", "vehicle", as well as semantically related words like "sedan", "driving", "highway", and "motor." As opposed to other linguistic approaches that are based on lexical semantics, PLSA does not need a language-specific (or even domain-specific) thesaurus or dictionary, but *learns* directly from the unstructured content. This has several key advantages: first, it ensures that PLSA is applicable to any language, as long as the language can be tokenized. Second, the extracted concepts are specific to the given document repository and are adapted to the language, technical terms, and specific jargon. Obviously, rebuilding similar thesauri by hand for each domain would be prohibitively expensive and time-consuming. Third, PLSA also learns a "numerical" model, where each word has some probability to occur in a certain concept. This allows the Axcelerate System to quantify the relationships between words.

Categorization

Recommind's CORE^{TM} (Context Optimized Relevancy Engine) Platform, on which the Axcelerate System runs, has categorization built deeply into the software indexing layer. This allows the user to interactively tag documents, or use existing meta-data tagging, and then use any meta-data tag as a seed set for categorization training. The resulting classifier is then applied against the complete corpus of documents. Both steps are tightly integrated into the indexed document representation, allowing for rapid training cycles that typically range from a few seconds to a few minutes for multi-million document corpora.

Recommind uses a variety of algorithms for classification and to build a statistical model of example documents. This model categorizes new documents as belonging or not belonging to the category of interest.

Ranking

Documents are mapped into a high-dimensional feature space and we compute a maximally separating hyper plane defining the decision boundary between documents inside the category and outside the category.

The distance to this decision hyper plane is used as a ranking function, i.e. the further away from the category boundary the higher the score for a document.

Probability Score Computation

In order to convert the distance score (which can be arbitrarily large) into a probability estimate on likelihood of being within the category, we again use the set of held-out test data. Using the distance function as a one-dimensional projection, we then use a simple logistic regression as a probability estimator. Essentially, this uses a sigmoid function to transform the distance score. The actual algorithm used here is a maximum likelihood approach to ensure that the transformation provides an actual probability estimate.

RESULTS

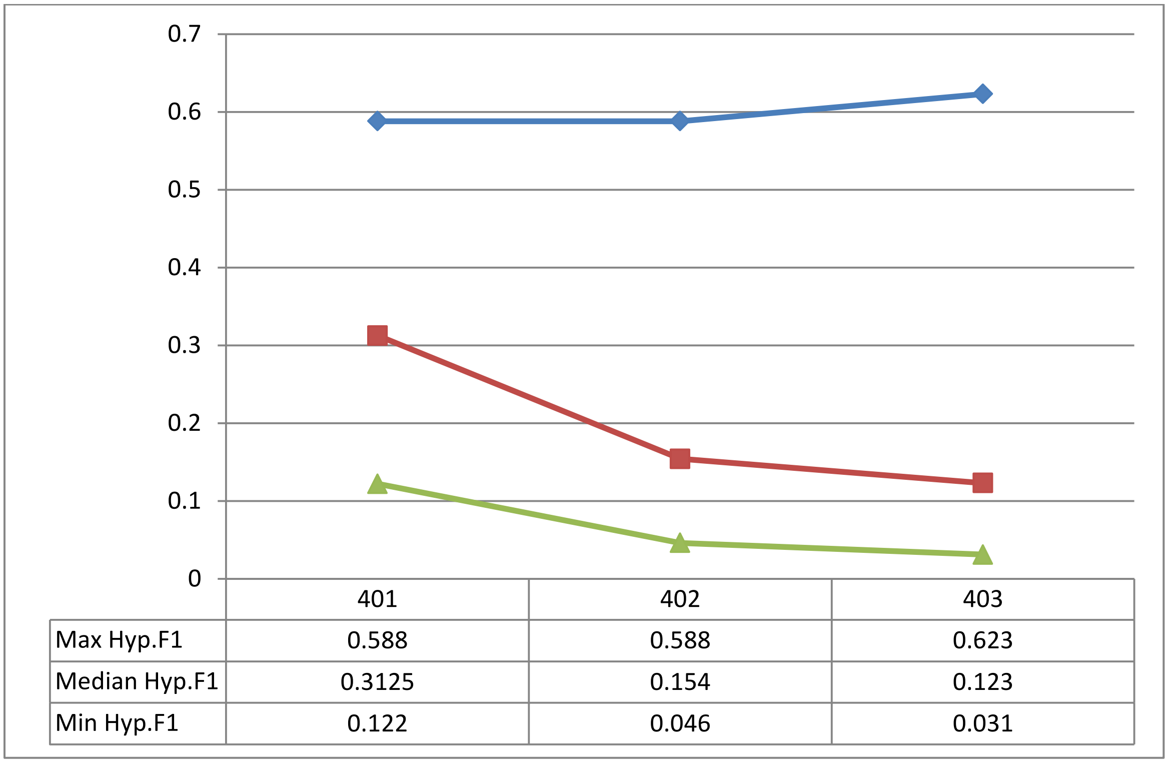
For its participation, the Recommind team produced two runs to TREC for evaluation: recommind03T and recommind04T. The primary difference between the runs was the method with which the responsiveness score was calculated. The responsiveness score in the recommind03T run was calculated by averaging the highest level review determination with the computer generated probability score. For example, if a document was coded at first and second level review, only the second level review determination and computer generated probability score would be taken into consideration when calculating the responsiveness score. The responsiveness score in the recommind04T run was calculated by averaging all available human review determinations along with the computer generated probability score. The method used in recommind03T was deemed to be more reliable and will be the focus of analysis.

Of the approximately 670,000 documents in the dataset only 38,000 went through human review, or 5.7% of the entire corpus.

TREC provided the best, median and worst hypothetical F1 scores per topic as a measure of success. Recommind achieved the highest hypothetical F1 scores for all three Topics (see Table 1).

Topic	Best	Median	Worst	Recommind
401	58.8%	31.2%	12.2%	58.8%
402	58.8%	15.4%	4.6%	58.8%
403	62.3%	12.3%	3.1%	62.3%

Table 1: Best, median and worst hypothetical F1 scores per topic in the Final Run.



	401	402	403
Max Hyp.F1	0.588	0.588	0.623
Median Hyp.F1	0.3125	0.154	0.123
Min Hyp.F1	0.122	0.046	0.031

Graph 1: Best, median and worst hypothetical F1 scores per topic in the Final Run.

Efficiency Gains

Efficiency Gain is a ration that looks at how many more documents would need to be reviewed in a alternative system to acheive the same level of recall as Recommind. For example, an Efficiency Gain of 10x means you would have to review 10x the amount of documents. This would result in a 10x review cost increase as the Efficency Gain is proporionate to review costs.

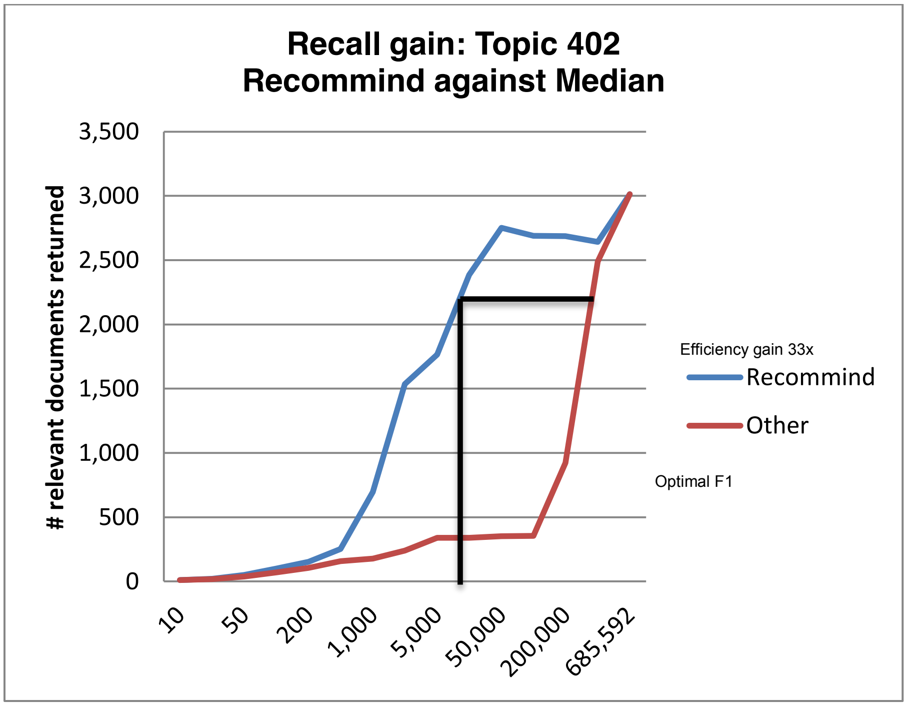
Steps taken to calcuate TREC Efficency Gains are:

- Extract scores from Final Run

- Remove all systems without results for all three topics
- Determine optimal cutoff point to achieve maximal F1
- Record number of relevant docs identified at that level
- Look at other participants and find point, where they would return the same number of relevant documents
- Compute the ratio between these two

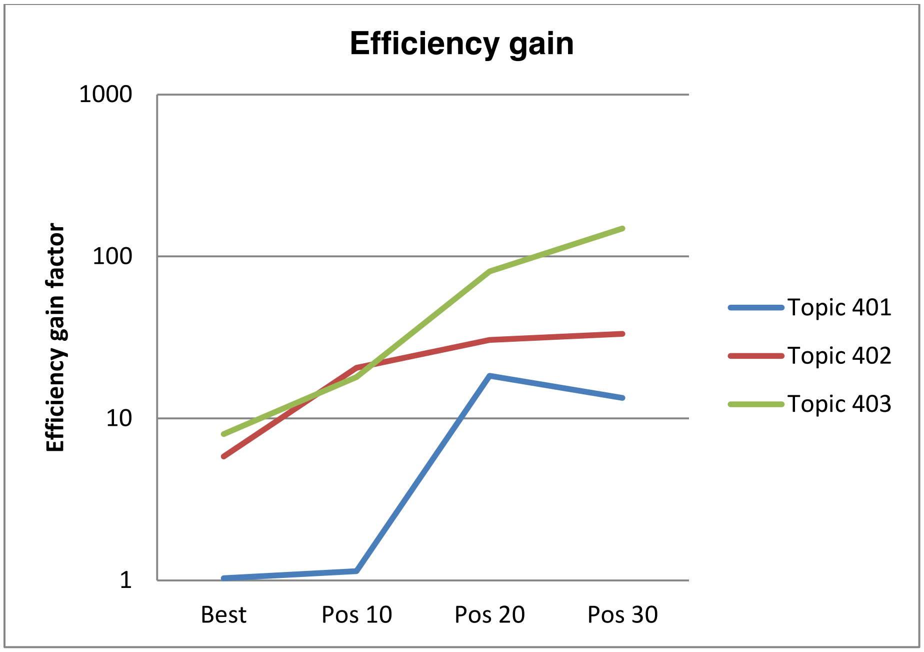
In addition to achieving consistently high scores for the Topics, Recommind also delivered significant gains in efficiency. On average, the Recommind solution boosted document review efficiency by a factor of 9.8x to 50x.

Graph 2 below illustrates the efficiency gains of Recommind as compared to the median for Topic 402.



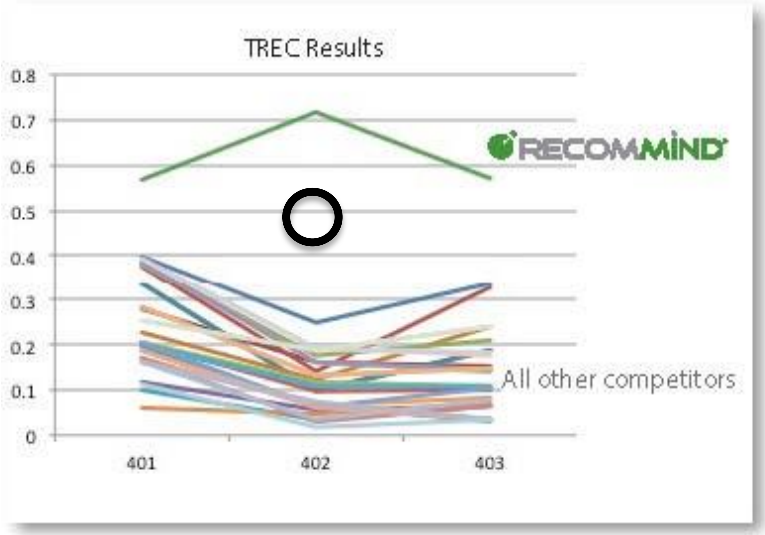
Graph 2: Efficiency gain for Topic 402

Graph 3 below illustrates Efficiency Gains for all three topics. Recommind was 9.8x more efficient than the 2nd place participants (average across the 3 Topics).



Graph 3: Average efficiency gains

Graph 4 below shows Recommind achieved the highest hypothetical F1 scores by a wide margin.



Graph 4: Hypothetical F1 scores across runs

CONCLUSION

The 2011 Task examined the responsiveness of the 685,000-document Enron data set to three Topics defined by the TREC Leadership and interpreted by TAs for each Topic. Utilizing its patented Predictive Coding process, the Axcelerate System received hypothetical F1 scores of 58.8%, 58.8% and 62.3% for the three Topics, respectively, after having “reviewed” (i.e. with human reviewers) in aggregate 5.7% of the entire corpus.

The methodology employed by Recommind was a key to the solution’s positive results. Recommind’s advanced analytics proved to be essential to obtaining high accuracy document reviews, particularly in the Early Review stage. This higher level of accuracy translated into efficiency gains of 9.8x to 50x over other TREC participants.

Predictive Coding includes a defensible workflow process that is a key factor in successful automated review and coding. By making a computer-generated judgment about the relevance, responsiveness, and/or privileged nature of each document, Predictive Coding allows Recommind to dramatically expedite the actual review process while concurrently improving accuracy and lowering the risk of missing key documents.

As expected, the Axcelerate System’s results in the final run were the highest of any participant for all three Topics. The Axcelerate System’s results in the 2011 Task clearly show that a) the Axcelerate System’s Predictive Coding process is a far superior approach than traditional linear review, and b) the Axcelerate System’s Predictive Coding process is the most effective form of semi-automated review methodology – and therefore, review methodology overall.