Clearwell Systems at TREC 2009 Legal Interactive

Venkat Rangan
Clearwell Systems, Inc.
venkat.rangan@clearwellsystems.com

Maojin Jiang
Clearwell Systems Inc.
maojin@clearwellsystems.com

ABSTRACT

The TREC Legal Track 2009 features an Interactive Task that is designed to replicate real-world challenges in producing a collection of responsive documents from a large collection of documents. The task required us to produce responsive documents from any of the seven topics, which are production requests. Clearwell Systems incorporated novel methods for producing a responsive collection using a combination of automated sampling, evaluation of the samples, and using the samples as input into a blind relevance feedback engine. The algorithms applied use an automatic correlation covariance matrix for automatic evaluation of the samples and, using the correlation coefficient, determine whether the process of blind feedback converges to a highly correlated set of responsive documents. The number of iterations of sampling, the K-value for blind feedback, along with the final convergence threshold are monitored. The F-measure results of this are compared across the three different Interactive Topics that Clearwell participated in, for discussions.

1. INTRODUCTION

In TREC 2009 Legal Track, Clearwell participated in the Interactive Task and worked on in total three topic requests, including topic 201 (prepay transaction), topic 202 (FAS 140/125 compliance) and topic 205 (energy load).

The Interactive Task is motivated by the real world litigation process and aims at modeling the real-world conditions in which e-discovery is pursued by law firms or other e-discovery companies [6]. From the previous Legal Track task and the current e-discovery practices, such a legal search embodied in the Interactive Task is an iterative process that requires collaboration among members of a team consisting of people from different areas [4, 7], a well-thought-out process [1, 5] and lastly but not the least importantly a powerful software system that allows and automates such collaboration and process as much as possible, in addition to providing state-of-the-art search techniques.

As a result, success of such an e-discovery task does not merely rely on advanced search techniques. It also calls for well-planned and close collaboration of searchers, especially with senior litigators, which are the roles that a Topic Authority (TA) plays in the Interactive Task. It is a senior litigator, or a TA, who defines the scope of responsiveness that each team as a searcher should have to replicate at the time when it performs searches to produce relevant documents to meet a topic request.

Thus, it is our interest both to explore state-of-the-art search techniques and to leverage the advanced case management function of Clearwell e-discovery systems for the Legal Interactive Task. In this paper, we will address our work in both directions.

The remainder of this paper is organized as follows. In the next section, we give a brief description of the data set used for the Interactive Task. In section 3, we introduce the Clearwell e-discovery system that we used throughout the execution of the three interactive tasks. Then, in section 4, we address our approaches to indexing, search, advanced analysis, case management, review and production and TA communication, after which we give our results in section 5. Lastly, we give the conclusion and discuss possible future work in section 6.

2. DATA COLLECTION

This year's test collection for the 2009 Interactive Task is Enron Collection, which consists of 569,034 unique email messages, in the native format of Microsoft '.msg' files, together with attachments, which are of a variety of file types, such as PDF files and JPEG images, etc. Including attachments and duplicate messages, it amounts to over two million documents. The collection also provides these email messages in plain text version, which contains only extracted text from messages and attachments. This collection is generated by Clearwell Systems from the collection that was originally released by the Federal Energy Regulatory Commission (FERC) and acquired by Clearwell from Aspen Systems. Further cleanup process was performed by University of Maryland.

Together with the collection are a mock complaint and a list of seven topic requests, out of which we were assigned topic 201, 202 and 205. Note that in the mock complaint, the imaginary major defendant is named Volteron Corp.

3. *ESA* – CLEARWELL E-DISCOVERY PLATFORM

Throughout the execution of the three topic request production, we used the Clearwell enterprise-class e-discovery platform to index and process the messages and attachments, perform searches, manage collection and search results as cases, generate samples of search results for review and assessment, and finally export search results for production. In addition, the sampling engine of the platform was enhanced to automatically sample both the retrieved and un-retrieved collections, and use the samples to measure dispersion of the sampled documents from the mean vector of the document feature vectors of the top N results from the original search. If the dispersion was beyond a certain established threshold, the samples are analyzed to automatically identify additional salient search terms for the next iteration of search. We then observed the convergence of this iterative search. We also evaluated a human judgment of samples and compared the results from the manual search query augmentation against the automated search query augmentation.

In this section, we discuss the key features of the Clearwell e-discovery system that are used for the Interactive Task. In the remaining part of this paper, we will use ***ESA*** to refer to this system.

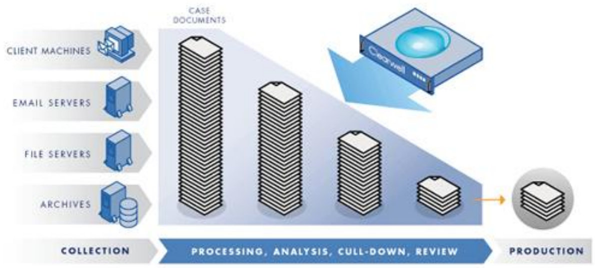


Figure 1. Clearwell E-Discovery Platform.

As depicted in Figure 1, ***ESA*** unifies the following key capabilities in its powerful case management, including *processing*, *analysis*, *search and cull-down*, and *review and production*, running as a Web application, easily accessible from a modern Web browser such as Internet Explorer.

Case management allows collaboration among multiple users to work on the same or different projects within or across multiple cases. In a case, different data sources can be indexed and analyzed for various e-discovery tasks. Users can be assigned different roles to access the complete or different parts of the case data, performing searches, reviewing, tagging and finally producing desired data for e-discovery production.

Processing provides capability to identify, index and analyze over 400 different document types. Particularly, it is able to read emails to identify and extract various regions, such as sender, recipient, subject and attachments, which facilitate various analysis functions. In addition, it also applies linguistic techniques such as sentence structure, part of speech information, word sense disambiguation to detect and label noun phrases in textual data, using commercial application from Basis Technologies[1]. As we know, an inherent property of email is its high redundancy in that multiple copies of the same message exist due to the fact that the same message often has more than one recipient. To handle this in order to improve the quality of search results, ***ESA*** also provides intelligent de-duplication step to detect such redundancy.

Analysis adopts proprietary, patent-pending algorithms known as “Dynamic Content Analysis^{TM}” to create the “Clearwell Master Index^{TM}”, which is much more sophisticated and powerful than a traditional full-text index. A message stream often weaves naturally into discussion threads, which group together the initial message, replies, carbon copies, forwards and near duplicates. By tracing a thread, one can quickly identify all the participants, and determine who knew what and when, which proves invaluable when performing early case assessments. Another analysis feature is topic clustering, which automatically groups messages based on topics they discuss by incorporating a text clustering algorithm derived from K-means hard clustering [8]. One last important analytic feature is to clean up, merge and optimize various data structures previously generated so that queries can be efficiently executed, also consisting of three types of analyses, namely 1) ‘people analytics’, which allows a user to access a list of top custodians for a search or monitor communications between regulated and non-regulated divisions within a company; 2) ‘file analytics’, which allows investigators to easily determine everyone who possesses or has sent or received a file of interest and allows reviewers to review a file once instead of multiple times; and 3) ‘term analytics’, which analyzes noun phrases to help users uncover secret project names and code words that may be relevant to a case or investigation by applying Natural Language Processing techniques (NLP).

Search and cull-down offers a suite of state-of-the-art search techniques through a user-friendly interface. Especially tailored to email search, it supports searches based on senders, recipients, direction, subject, attachment names and types, and date range, etc. As we know, emails often demonstrate a free writing style, which means senders who write messages do not pay too much attention to spelling or grammar. To cope with this, ***ESA*** supports wildcard, fuzzy and proximity searches as well. Another useful search feature is facilities for interactive query expansion (IQE) [9,10], packaged as a ‘search preview’ (an illustration of this feature is shown in Figure 3, Section 4.3). This provides visibility into matching keyword variations for wildcard and stemming searches prior to running a search. Users can see all variations of keywords used in search with their frequencies in the case data, which makes it possible for users to selectively include relevant variations or exclude false positive variations from their search query, thereby adding relevant documents and removing irrelevant documents from search results. In order to assess the quality of a search, a popular method is to make a sample of search results for assessment. This can be done within ***ESA*** by either manually selecting documents or by automatic and random selection, at a user's discretion.

Review and production provides multiple viewing modes giving users the flexibility to view header, snippet or document detail. In addition to linear review of individual documents, ***ESA*** also allows users to view documents as a group based on discussion threads or topic clusters, significantly increasing the review throughput. In addition, ***ESA*** allows case administrators to bulk tag documents into review sets and assign them to an individual for review. Case reviewers can review documents in their favorite reviewing modes and easily view and update tags for documents. Lastly, ***ESA*** allows results to be produced in multiple production formats (e.g. CSV or XML) that may meet different needs for different cases or tasks.

4. METHODOLOGY AND TASK EXECUTION

This section describes the methods and the process we used during the execution of three interactive tasks. A diagram of the execution is depicted in the Figure 2. Simply speaking, a searcher constructs a search based on his or her understanding of the information needed behind a topic request, performs the search, then communicates with a TA about the effectiveness of the current search (e.g. by asking a TA to judge responsiveness of a document, etc.). In addition, search results are randomly sampled for internal assessment. Then, by reviewing both the TA’s feedback and judgment of the current search and the internal assessment result, the searcher will adjust or correct his or her

[1] Basis Technologies, Rosette Base Linguistics for English, http://www.basistech.com/base-linguistics/english/

understanding of the topic request and construct a new search, which essentially repeats the above steps.

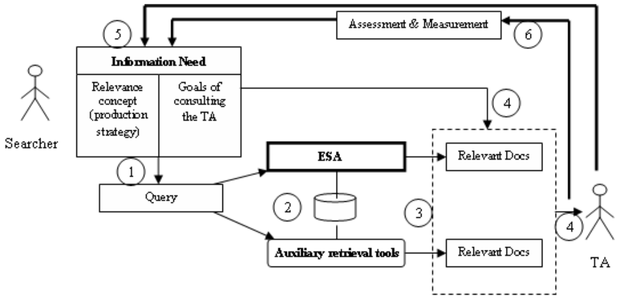


Figure 2. Interactive Task Execution Diagram

4.1 Iterations

In our execution of the tasks, six iterations were performed and tracked. Most of the iterations had a duration of one week. The first iteration spanned over a month after the task guideline and court complaint were released by TREC and until an orientation call was given for a topic. The last iteration lasted over one week.

4.2 Processing

All messages and attachments in their native formats were processed as follows. No word was removed (this enables us to do some pattern-based search; for example, we can search and find out noun phases following 'such as' in the collection). Terms were stemmed after tokenization. Further, all terms were lowercased and indexed. In addition, different regions of an email were identified and indexed separately, such as 'subject', 'sender', 'recipient', timestamps, attachment names, etc., which facilitated search by these regions later. Thanks to the capability of supporting over 400 different file types in ***ESA*** in native formats, an individual attachment was detected with its original file name in each message and its textual content was able to be correctly parsed and tokenized, which maintained the integrity of a parent message and its attachment(s).

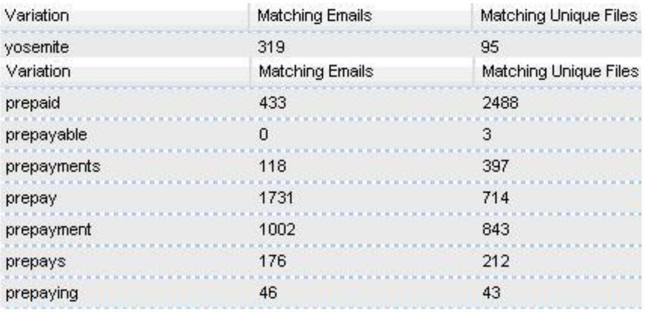
Beyond simple tokenization that is often done on unstructured plain text and indexing, we also performed three types of post-processing 'analyses' as discussed in section 3 previously. By these steps, structures of emails were analyzed and retained such that the related messages that are replies or forwards of the same messages were grouped together as a discussion thread. The messages were also grouped by participants so that we can easily and efficiently find out who sent what messages to whom and, from the timestamps, when they were sent. Duplicate messages were detected and labeled. Topics were discovered by applying NLP algorithms. Finally, data storages were optimized for efficient searches.

4.3 Search

First, we take advantage of the powerful yet flexible Boolean search features provided in ***ESA***. This year's submission is based on "document-level" responsiveness. That is, each individual attachment and each parent message are considered separate individual documents. So, to produce any relevant document, we need to restrict a search to either the parent message or to the attachments to locate the occurrence of relevant information. In ***ESA***, we can easily achieve this by searching by regions, including 'subject line', 'sender', 'recipient', 'attachment file name', 'quoted text in a message', 'message body without quoted text' and attachments, among other things.

Another important feature we used in ***ESA*** is called "search preview" as aforementioned. Using it, we can determine if we need to search by using variations of a search word in order to get an estimate of how many unique emails and how many unique files may contain the word and its variations. This is especially useful in handling free writing styles in email settings since it is not rare that neither grammar nor spelling is strictly followed in writing messages. For example, Figure 3 shows the variations of the words 'yosemite' and 'prepay'.

As we know, a message is often forwarded or replied to more than once. Such follow-up messages together with the original message form a natural discussion thread. Although sometimes arguable, it is often true that if a message is relevant, its follow-up messages can be also deemed relevant. As a result, in our work, once we found a relevant parent message, we also included the messages in the same discussion as relevant. In the future, we will pursue discriminative power at a finer granularity based on this. For example, if a follow-up message does not quote the original message, and does not contain any new text that is relevant, the message will not be considered relevant.



Variation	Matching Emails	Matching Unique Files
yosemite	319	95

Variation	Matching Emails	Matching Unique Files
prepaid	433	2488
prepayable	0	3
prepayments	118	397
prepay	1731	714
prepayment	1002	843
prepays	176	212
prepaying	46	43

Figure 3. Search Preview of 'yosemite' and 'prepay'.

In e-discovery, especially in searching for documents for civil litigation use in courts, the information need is often so complex that it is hard to describe as keyword searches at first sight. For example, topic 202 request asks searchers to find "*all documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in transactions that the Company characterized as compliant with FAS 140 (or its predecessor FAS 125)*". Although some documents will mention FAS 140 or FAS 125, one would expect that most responsive documents will not mention these words at all. As a result, to find these 'hard' documents that talk about either FAS 140 (or FAS 125) compliance or in-compliance but do not mention any of these words such as 'FAS', '140', '125', or 'financial accounting statement', we need to understand the concepts conveyed in FAS 140 or FAS 125 and their relationships, in order to find these 'hard' documents.

To deal with such difficulties, two auxiliary search methods are used. One is collocation (for both topic 201 and 202) and the other is concept search (solely for topic 202).

Based on collocation, we observed that even when we do not have a profound understanding of concepts to be searched, we can take advantage of what is collocated with known entities of given concepts to find new entities. In essence, what we used is based on bootstrapping information extraction that was proposed in the ‘KnowItAll’ system [3]. In contrast to the proposal, in our work,we only performed one iteration by searching for certain patterns. For example, based on known ‘Hawaii 125’ deals, we found other deals that might be highly relevant to topic 202 by searching for noun phrases close to “Hawaii 125 series”, including ‘*McGarret C, D, G, F(Riva)(New) H(Braveheart) (New) and Danno B(Alchemy)*’. For another instance, based on known financing vehicles, such as ‘Osprey’, we found other similar vehicles by searching for the following patterns, “financing vehicles include NP (, NPs)” (or “such as NP (, NPs)”) appears within a small context window of ‘Osprey’.

Concept search is leverages the fact that semantically correlated terms often occur together. The application of such correlation is first adopted in Latent Semantic Analysis (LSA or LSI standing for Latent Semantic Indexing) [2]. To capture the concepts and their relationships conveyed in FAS 140 or 125, we also used FAS 140 summary and each individual paragraph (including paragraphs 2, 5, 6, 8, 9c, 17d, 17f, 27, 35 and 36, as mentioned by the topic 202 TA) in a concept search to find documents that might mention the compliance or in-compliance with that specific requirement in the summary and each paragraph in the statement. After giving samples of these concept search results to the TA and collecting his judgments, we only used the top relevant results found by using paragraphs 5, 6, 8, 9c, 27 and 36 in the end.

We also performed blind relevance feedback (BRF). By returning the top salient terms in the top relevant documents, we did find some clues which may lead to finding more responsive documents; for example, one such term found in this way is *CSFB* for topic 201, which refers to ‘Credit Suisse First Boston’, involved in some circular prepays with Volteron. But overall, these terms were already known to us through other methods and the number of these useful terms found is small. So, we did not use BRF as a means to find more responsive documents.

In the last iteration, by summarizing the effective searches from previous iterations, we ended up with a long list of Boolean queries for each topic (for topic 202, in addition to Boolean queries, we also have concept searches).

4.4 Case Management: Sampling, Reviewing and Production

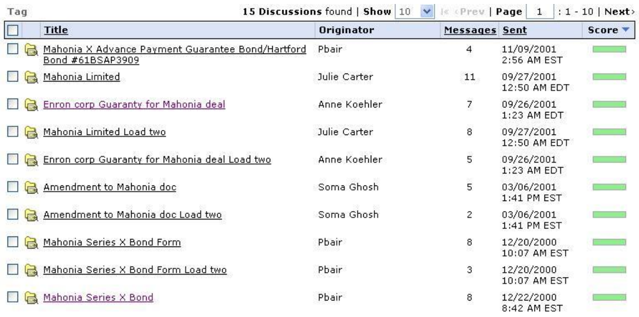
As mentioned in the Introduction section, an interactive task naturally progresses into an iterative process, which demands management of search and search results from different searchers and in different stages, allowing for sampling, collaborated review and production.

With this in mind, we created a case in ***ESA*** for the collection in the native formats. The case was carefully named with a description to indicate what it is about and the location of its data source. In the case, exactly one project was created for each topic for the duration of each iteration. Within a project, searchers were allowed to create tags to label different search methods. After a search was done, the documents found were labeled with the tag of the corresponding search used.

To review the effectiveness of each search method, the following process was followed to allow reviewers to review a sample set of results: first, an administrator created a list of users for search result reviewing and assessment, and for each user, a project was also created; next, the administrator created a randomly sampled set from search results to be assessed and added it to each reviewer’s project; the administrator also created a global multiple-valued tag to allow each reviewer to tag a document under review as “Responsive”, “Not Responsive” or “Not Sure”. In this way, by collecting different values of the global tag, we were able to know how many results were assessed as responsive, not responsive or not sure.

As discussed earlier, messages in the collection were processed to recover discussion threads. ***ESA*** provides a convenient visualization of messages in a discussion thread for reviewing. For example, Figure 4 (a) illustrates the first 10 out of 15 discussions detected from all messages with ‘mahonia’ in their subject lines. Figure 4 (b) shows an expanded structure of four participating messages by selecting the discussion ‘Mahonia Series X Bond’. In the figure, the participating messages are labeled by their senders and are shown in the left panel. The content of the currently selected message is shown in the right panel, which is further divided into two panes: the lower one displays the forwarded text and the upper displays the new text. Note that the search term ‘mahonia’ is also highlighted.

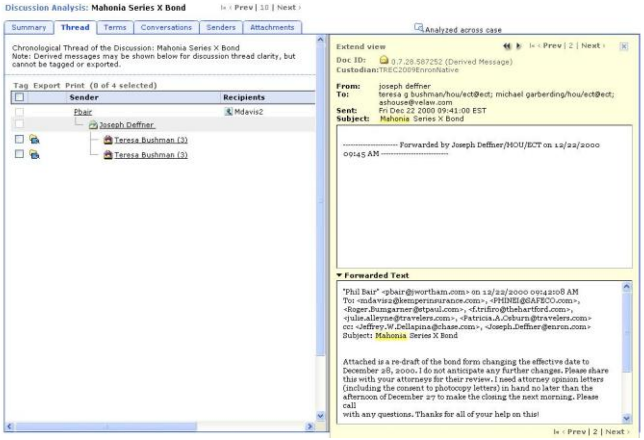
Having documents tagged with different names, ***ESA*** allows a user to combine them in various ways to obtain a desired merged set, e.g. including documents labeled with one specific tag or subtracting documents labeled with another one from the final set. By managing search and search results in this way, good searches and positive results can be carried over from a previous iteration to a latter one while allowing updates.



Tag
15 Discussions found | Show 10 | |« ‹Prev | Page 1 : 1 - 10 | Next›

□	Title	Originator	Messages	Sent	Score
□	Mahonia X Advance Payment Guarantee Bond/Hartford Bond #61BSAP3909	Pbair	4	11/09/2001 2:56 AM EST	🟩
□	Mahonia Limited	Julie Carter	11	09/27/2001 12:50 AM EDT	🟩
□	Enron corp Guaranty for Mahonia deal	Anne Koehler	7	09/26/2001 1:23 AM EDT	🟩
□	Mahonia Limited Load two	Julie Carter	8	09/27/2001 12:50 AM EDT	🟩
□	Enron corp Guaranty for Mahonia deal Load two	Anne Koehler	5	09/26/2001 1:23 AM EDT	🟩
□	Amendment to Mahonia doc	Soma Ghosh	5	03/06/2001 1:41 PM EST	🟩
□	Amendment to Mahonia doc Load two	Soma Ghosh	2	03/06/2001 1:41 PM EST	🟩
□	Mahonia Series X Bond Form	Pbair	8	12/20/2000 10:07 AM EST	🟩
□	Mahonia Series X Bond Form Load two	Pbair	3	12/20/2000 10:07 AM EST	🟩
□	Mahonia Series X Bond	Pbair	8	12/22/2000 8:42 AM EST	🟩

(a)



Chronological Thread of the Discussion: Mahonia Series X Bond
Note: Derived messages may be shown below for discussion thread clarity, but cannot be tagged or exported.

□	Sender	Recipients
□	Pbair	Mdavis2
□	Joseph Deffner	
□	Teresa Bushman (3)	
□	Teresa Bushman (3)	

Doc ID: 0.7.28.587252 (Derived Message)
Custodian:TREC2009EnronNative

From: joseph deffner
To: teresa g bushman/hou/ect@ect; michael garberding/hou/ect@ect; ashouse@velaw.com
Sent: Fri Dec 22 2000 09:41:00 EST
Subject: Mahonia Series X Bond

---------------------- Forwarded by Joseph Deffner/HOU/ECT on 12/22/2000 09:45 AM ---------------------------

"Phil Bair" <pbair@jwortham.com> on 12/22/2000 09:42:08 AM
To: <mdavis2@kemperinsurance.com>, <PHINEI@SAFECO.com>, <Roger.Bumgarner@stpaul.com>, <f.trifiro@thehartford.com>, <julie.alleyne@travelers.com>, <Patricia.A.Osburn@travelers.com>
cc: <Jeffrey.W.Dellapina@chase.com>, <Joseph.Deffner@enron.com>
Subject: Mahonia Series X Bond

Attached is a re-draft of the bond form changing the effective date to December 28, 2000. I do not anticipate any further changes. Please share this with your attorneys for their review. I need attorney opinion letters (including the consent to photocopy letters) in hand no later than the afternoon of December 27 to make the closing the next morning. Please call
with any questions. Thanks for all of your help on this!

(b)

Figure 4. Discussion Visualization Tool

In the last iteration, similarly, a batch of searches was performed, documents found were tagged and a final set was generated by merging results of different searches. Lastly, for production, we used ***ESA*** to export a list of the final set which shows the original document ID associated with each result in the collection by including those messages in the same discussion threads as well.

4.5 Automatic Sampling

We also implemented an automatic sampling evaluation system that enabled us to evaluate convergence of the search process. Given a set of searches that were validated by human review, our retrieval process sampled the non-retrieved collection, and evaluated each individual document's similarity to other documents in the retrieved set. This similarity measurement was based on Noun Phrases extracted from the text of emails and attachments. Previous studies [11,12,13] indicate that the information content from noun phrases is sufficient to describe the features of a document. We compute a scored feature vector for each document, based on its frequency of occurrence in various regions of text of emails. Also, we selected the top 20 noun phrases, as the information gain from the lower ranked noun phrases is small. This feature vector from the sample document is verified for similarity using the cosine distance formula, to identify whether the document from the sample is close to any of the documents of the retrieved set. The number of documents that are part of the non-retrieved set that is greater than a threshold cutoff in similarity represents missed documents that would reduce the recall rate. Given the overall goal of achieving a high recall, we then analyzed the documents with high similarity for additional noun phrases that must be used to for the next iteration of the search. This constitutes a single iteration of search relevance feedback.

To evaluate convergence of subsequent iterations, we measured the number of documents that were in the missed pool. We expect convergence if the information gain in the new feedback loop is less than previous iterations, and if the additional documents identified are below a certain threshold document count.

We represent each document by its feature vector, V= $v_{i=0,N}$ where each v_i represents a noun phrase in a score-ordered list of noun phrases extracted from that document. Each sampled document has a feature vector, S_v which is then measured for similarity against a featured vector that represents the entire retrieved set. The retrieved set feature vector similarity was measured using two different methods: i) merged feature vector evaluation and ii) document-by-document feature vector similarity evaluation.

The merged feature vector comparison first combines the top score-ordered documents from the retrieved set and merges their feature vectors. For this study, we placed a *k* cutoff value of 2000 retrieved documents, whose feature vectors are then merged per the following formula.

Each feature vector entry, v_i is represented as a tuple $\langle t_i, s_i \rangle$, where t_i is the raw term frequency and s_i is the score for the term. The merging of feature vectors into a combined feature vector retains the term frequency of all the vectors and normalizes the score by the total number of terms in the feature vector.

The similarity of the sampled document and the merged feature vector is based on the cosine measurement, computed using the following.

$$Similarity = \frac{\sum_{i=0}^{N} V_i * C_i}{\sqrt{\sum_{i=0}^{N} V_i * V_i} * \sqrt{\sum_{i=0}^{M} C_i * C_i}}$$

This assumes that the document's vector V_i has [0, N] noun phrases and the merged feature vector has [0, M] words each with frequency c_i . For the specific noun phrase t_i , the corresponding word's frequency in the cluster feature vector is c_k . If the document feature word does not appear in the cluster feature vector, this word contributes zero to the dot product.

For document-by-document feature vector evaluation, we compute each pair-wise similarity and note the number of pairs where the similarity exceeds a certain threshold.

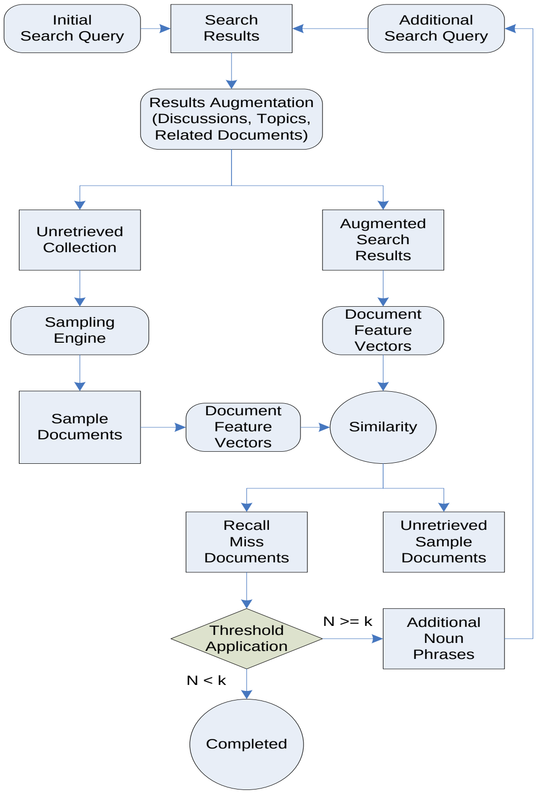


Figure 5. Automatic Sampling Iterations

Sampling evaluations were designed to constrain the error and confidence levels to standard practice. In our evaluations, we used a confidence measure of 95%, so that the sampling error is within ±5% of the estimated value. This resulted in us evaluating 1537 sample documents for a coverage of one sigma around the mean of the distribution.

4.6 TA communication

In real world litigation, a leading attorney defines the scope of document production. In the interactive task, for each topic, a Topic Authority (TA) plays the same role.

Clearly, the scope of the information needed for each topic is beyond the surface meaning of the topic description. As a result, to draw a clear picture of what is requested to produce, all three TAs gave an orientation call at the beginning of the task execution and we participated in all of them.

These initial orientations proved to be very helpful to correct our initial understanding of topic requests. For example, by description, topic 205 seemingly asks for all documents that talk about volume(s) or geographic location(s) of energy loads. As a result, before we attended the orientation, we tended to produce any document in which some energy products of Volteron Corp. (the imaginary defendant) are discussed in volume, e.g. 10 MMBTU natural gas or 10 MWh electricity[2]. However, this is incorrect. In the orientation, the TA strengthens that it is the overlap of paragraph 21e of the mock complaint with the the description of the topic that defines the scope of production. From paragraph 21e and other paragraphs in the mock complaint, we know that Volteron participated in manipulation of the U.S. energy market in which false volume(s) and geographic location(s) are involved to illegally raise prices of their energy products. With this new restriction, for example, it immediately put the production under attack, since the production contains a document that simply summarizes the volume of natural gas provided for a European country in a specific year. As a result, this reminds us (as a searcher) that we should not isolate a topic request from the complaint whenever we determine whether a document is responsive or not. In addition to putting a restriction on documents speaking of volumes, the orientation also revealed that we needed to produce those messages and attachments that relate to Volteron's manipulation in various ways, directly or indirectly. One such example is that Volteron sells a certain volume of electricity to another state and then immediately buys it back to give a false impression of a lack of energy in the home state and to ask for a congestion fee from the state's regulatory organization.

We communicated with TAs of topic 201 and 202 mostly by emails and with TA of topic 205 mostly by phone. In the first contact, we exchanged thoughts for both logistics and some general production questions, such as whether a follow-up message should be produced given that the message it follows is responsive. Then, throughout the execution of a task, we gave each TA our findings with our rationale and asked for his judgments. Based on his feedback, we made adjustment to our understandings and search methods.

In the end, we ended up with spending 4 hours, 5 hours and 8 hours in consulting TAs of topic 201, 202 and 205 respectively.

Although the interaction with three TAs is helpful, we found some confusion regarding the overlap of the role of the interactive task teams and that of a TA. Here is why we have such confusion. For example, in the topic 205 task, participating teams are asked to find all 'circular prepay transactions' in the collection. Our assumption is that the TA will help all teams to understand what a 'circular prepay transaction' means and then it is each team's task to find any such prepays appearing in the collection. However, right before the orientation call, the TA distributed a 192-page document that lists many details of 11 circular prepay transactions. In our opinion, we believe these 11 prepays, together with other circular prepays that might appear in the collection, are what participating teams should look for. So, any information related to such prepays should remain un-disclosed until the submission is closed and will then be used as facts to judge the final results of teams.

5. RESULTS

We submitted 16112, 20937 and 292540 results for topic 201, 202 and 205 respectively, including both messages and attachments, at document-level.

During generation of the final results, we found the following issue. In the collection, there is a list which shows the identifiers of unique messages, which are called master copies. Another list shows messages that are duplicates of each master copy. Ruled by TREC, only the master copy messages and their attachments will be used in the final evaluation of interactive tasks. In our work, we used all the messages and thus their attachments in the

[2] Natural gas and electricity are energy products of Volteron.

collection for all three topics and found responsive messages and their attachments listed as duplicates as well. However, many such duplicates are actually different from their master copies (or their attachments) in content (probably due to some de-dup errors during the time the duplicate list was prepared). As a result, they should be produced to reflect such differences. But unfortunately, they will be excluded from evaluation because they are deemed as duplicates by TREC.

Because of this, we made the following adjustment for these duplicate messages and attachments whose master copies were not produced in the final result for a topic: 1) take the identifier of each such message, replace it with the identifier of the master copy; and 2) take the identifier of the message part of such an attachment, replace it with the identifier of the master copy.

5.1 Overall Measurements

The following shows the evaluation results of our submission. An important goal that we established was to give preference to Recall at the expense of Precision.

Topic	Submitted Count	F1	Recall	Precision
201	16112	0.299	0.489	0.215
202	20937	0.619	0.579	0.664
205	292540	0.410	0.673	0.321

Table 1: Adjudicated Results

Our goal was to identify the highest number of responsive documents using an iterative procedure. In that context, we completed 18 iterations of various searches. Each iteration was driven by the following items for refinement.

a) Additional search terms suggested by the ESA Search Preview, which is based on surface features of keywords such as stemming, wildcards, fuzzy matches etc.

b) Additional Conceptual Matches.

c) Related Terms, as defined by other terms that co-occur with previous search terms.

d) Additional messages from Discussion Threads and email conversation connections.

e) Additional messages that enclosed the same or similar attachments.

f) Automatic sampling process to identify sampled messages that matched a significant number of documents. A threshold of 80% similarity was required for a document to be similar.

g) We also had an internal team of assessors evaluate a small sample of documents from the un-retrieved collection. Given that human review was expensive, we limited this to 50 documents per reviewer, with a total of 400 documents reviewed.

h) Small number of samples reviewed by Topic Authority.

The following progression of iterations illustrates the information gain for each iteration of results.

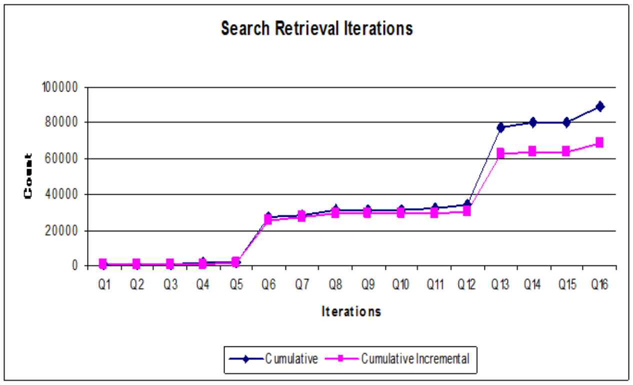


Figure 6: Retrieval Effectiveness

Note that some number of documents is duplicated during subsequent searches, and this causes additional new documents to be smaller resulting in less new gain in recall.

5.2 Sampling Distribution

An important consideration in determining a conclusion was whether there would likely be any new document gain from additional searches. While it is possible to manually review a sample of documents, we wanted to approach this using an automated sampling process, given the large number of iterations. An observation on manual review has been that each new iteration carries with it a new review cost, which is often substantial. One major innovation in our research is to identify the stopping point when we expect no new improvement in retrieval effectiveness for the cost.

To achieve this we sampled at 1537 samples (95% confidence for ±5% of error estimate) and identified whether new samples with high similarity added any new interesting search terms.

	Results	0.0	0.1	0.2	0.3	0.4	0.5	0.6
Q1	535	974	382	142	35	4	0	0
Q2	661	1042	397	62	36	0	0	0
Q3	255	1144	304	25	58	5	1	0
Q4	71	1292	195	34	10	3	2	1
Q5	606	939	326	211	60	1	0	0
Q6	24909	935	410	123	42	26	1	0
Q7	1685	926	517	77	11	3	2	1
Q8	2399	882	461	169	18	2	5	0
Q9	628	1000	477	46	10	4	0	0
Q10	26	1311	160	55	4	4	1	2

Q11	1032	988	513	31	5	0	0	0
Q12	1907	869	475	168	21	1	3	0
Q13	42524	845	595	90	7	0	0	0
Q14	3399	985	395	113	40	2	1	1
Q15	152	1017	360	131	29	0	0	0
Q16	8488	1027	444	52	9	3	1	1

Table 2: Topic 201 Sample Distribution

The above sample distribution illustrates the number of documents from the sample of un-retrieved documents that had a similarity to the merged feature vector of the top 2000 retrieved results. As can be seen, we see a general drop in sample match count at higher levels of similarity. Also, it was observed that at lower levels of similarity, commonly occurring terms such as "load", "gas", and "enron" tended to contribute similarity. On the higher similarity buckets, we found certain highly relevant terms that could be used for new searches. As an example, the term "Yosemite" was found in samples that matched against results from the term "Mahonia".

In addition to the distribution of samples, we measured individual matches between samples from the un-retrieved set against the retrieved documents. This is a measure of individual document-by-document matching of sample documents against retrieved documents. As can be seen, we found very few sample documents from the un-retrieved collection that matched documents in the retrieved collection.

Query	Results	Misses	Matching Misses	Miss Estimate
Q1	535	2	17	2750
Q6	24909	6	111	8251
Q10	26	2	2	2750
Q11	1032	1	1	1375
Q12	1907	1	8	1375
Q13	42524	6	14	8251
Q17	8488	1	2	1375

Table 3: Topic 201 Sample Misses

5.3 Discussion of Results

The TREC 2009 Legal Track exercise revealed certain process related items that appeared to impact the results. The Legal Track process involves an initial assessments of samples, where stratified samples are drawn from a population that had no team submissions, population that had a single team and multiple team submissions. These initial assessments are distributed to the participating teams for appeals of the assessments, which are then adjudicated by the Topic Authority. The appeal and adjudication phase has the potential to change the results from initial assessments in fairly substantial ways, when there are reversals on the initial appeals. One of the considerations is the rate of reversal of the initial appeals and its impact on adjudicated results.

Following are the rates of reversals, as indicated by the TREC 2009 Coordinators.

Topic	Samples	Appeals	Success	Reversal %
201	6956	497	464	93.4
202	7435	708	584	82.5
205	6367	967	932	96.4

Table 4: Topic Appeal Effectiveness

With a very large reversal rate, one consideration is its impact on assessments when not all the initial assessments were appealed. As an example, for those teams that appealed based on random selection of a subset of the samples, projecting the reversal into the unappealed population would in fact produce a more accurate reflection of that team's ability to retrieve relevant documents. Thus, the Legal Track Interactive Task changed from a pure information request into a review of first pass assessments exercise. Teams that invested review resources during this phase of the project benefited the greatest.

5.4 Impact of Appeal and Adjudication Phase on Clearwell Results

As a participating team, Clearwell chose to appeal only 2% of the sample assessments, and therefore did not fully benefit from the appeals and adjudication phase.

Topic	Appeals	NNN	R	Appeals Rate %
201	139	100	39	1.99
202	150	110	40	2.01
205	140	113	27	2.19

Table 5: Clearwell Appeal Rates

The impact of low appeal rate is shown in the change below, from an initial assessment of F1 estimate, which while improving from (0.088, 0.272 and 0.455) to (0.299, 0.619 and 0.434), the improvement was not as much as for other teams that chose to appeal the assessment in larger numbers. As an example, the remaining 558 appeals in Topic 202 changed the Team X results from 0.251 to 0.764, illustrating the impact of a more thorough review for generating appeals.

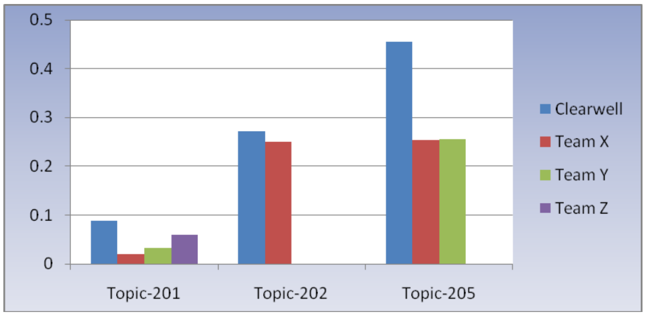


Figure 7: Clearwell Initial Assessment (F1)

The final adjudicated results compared with other participating teams is shown below. As noted, while all teams improved their estimated F measures, some teams benefited more than others, based on the extent of appeals.

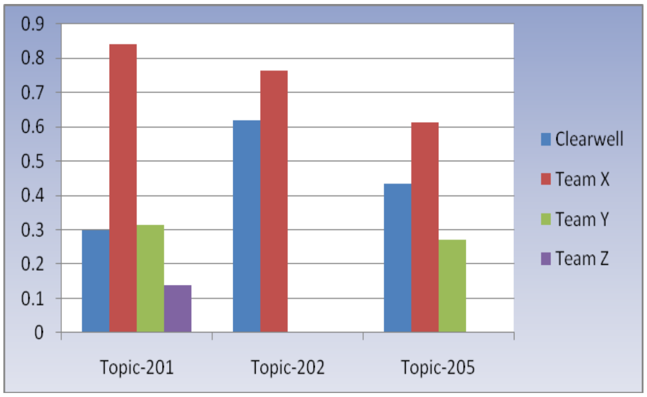


Figure 8: Clearwell Adjudicated Results (F1)

One potential way to manage this discrepancy is to limit the number of appeals each team can submit. Another way to extend the results from the appeal subset to the entire population. Assuming random selection of appeals, and the sample population reflecting a homogeneous mix, the following changes are anticipated.

Topic	Original F Estimate	Adjustment Ratio	Adjusted F Estimate
201	0.299	1.387	0.414
202	0.619	1.268	0.785
205	0.434	1.169	0.507

Table 6: F Estimate Adjustent

However, the overall impact of successful appeal of unappealed samples is impossible to predict. It is possible that a successful appeal from the rest of the unappealed population would reverse the R assessment of another team, causing their precision component to drop.

5.5 Topic 205 Anomalies

Clearwell noted that all 27 of their appeals to reverse Non-Responsive documents to Responsive were rejected. Upon reviewing the guidelines and the appeals, it is unclear why these appeals were rejected. Especially unclear is why identical document content, but near-duplicate of the appealed documents were assessed in one case as relevant, and in another case as not relevant.

Additionally, given the Topic 205 was a broad information request, and we were given guidance to consider the information request as broadly as possible, it came as a surprise to see the initial submission and appeals reverse this directive. This seems to indicate some high-level misunderstanding of the information request either by the Participating Team or the Topic Authority. Not having the ability to clarify this, especially in the context of very large reversals has the potential to leave the final assessments in question.

Another complication with Topic 205 was that this topic was about energy loads and manipulation of energy loads, but several thousand documents had a “Load Two” extraneous text added to email subject line. This text was to tag emails as a loading batch, but that addition complicated the search process.

6. CONCLUSIONS

In this paper, we reported the execution of three interactive tasks and results. We especially demonstrated how the Clearwell E-Discovery System can be fully explored to benefit e-discovery applications such as a search in civil litigation as simulated by an interactive task in Legal Track, from providing powerful and state-of-the-art search methods to managing the iterative search, review and production process executed and collaborated among human searchers. An important contribution is the incorporation of automatic sampling and using sampling as a way to supplement human review for determining the effectiveness of each iteration of search. Another technique we will explore is reference resolution. As we know, in messages, authors often use acronyms, code words (sometimes even coined words that cannot be found in any extant dictionary), or pronouns to refer to what was discussed in previous messages. If a message does not contain any explicit occurrence of relevance but does contain such references which point to relevant information elsewhere, we should produce the message as responsive. In addition, noticing this year’s interactive task will generate a collection of responsive documents for each topic, we also plan to investigate how supervised learning can be used to produce more accurate results by using this collection as training data. Lastly, it is easy to see that all emails form a social network naturally by connections (i.e. a person is a node and an edge connects the sender node to the recipient node(s)). Further, different overlay networks can be constructed by subject lines, by topics, by locations, or by time range, etc. In such social networks, each edge can be assigned a weight bearing certain semantics; for example, the similarity value by the topics of messages exchanged between peer nodes. One would expect that responsive messages might exhibit certain patterns in these social networks at a specific time point or over time. So, we also plan to do social network analysis to complement other methods in the interactive tasks in the future.

7. REFERENCES

[1] Jason R. Baron, Richard G. Braman, Kenneth J. Withers, Thomas Y. Allman, M. James Daley and George L. Paul 2007. The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery. The Sedona Conference Journal. Vol. 8 (Aug. 2007), 189-223.

[2] S. Deerwester, Susan Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41 (6): 391–407. 1990.

[3] Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2004. Web-scale information extraction in KnowItAll (preliminary results). In Proceedings of the 13th International Conference on the World Wide Web (WWW 2004).

[4] Christopher Hogan, Dan Brassil, Esq. Shana M. Rugani, Jennifer Reinhart, Misti Gerber and Teresa Jade. 2008. H5 at TREC 2008 Legal Interactive: User Modeling, Assessment & Measurement. TREC 2008.

[5] Jason Krause. 2009. In Search of the Perfect Search: A project closes in on a protocol to improve e-discovery results. ABA Journal.

[6] Douglas W. Oard, Bruce Hedin, Stephen Tominson and Jason R. Baron 2008. Overview of the TREC 2008 Legal Track TREC 2008.

[7] Zhen Yue, Jon Walker, Yi-Ling Lin and Daqing He 2008. Pitt@TREC08: An Initial Study of Collaborative Information Behavior in E-Discovery. TREC 2008.

[8] Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley. MR0405726 - ISBN 0-471-35645-X.

[9] D. Harman, Towards interactive query expansion, SIGIR '88: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval (1988), pp. 321-331.

[10] Bast, Holger; Majumdar, Debapriyo; Weber, Ingmar, Efficient interactive query expansion with CompleteSearch, CIKM'07 : Proceedings of the 2007 ACM Conference on Information and Knowledge Management.

[11] Matthew Chang, Chung Keung Poon, Using phrases as features in email classification, The Journal of Systems and Software, 82 (2009) 1036-1045, www.elsevier.com/locate/jss

[12] Yongli Liu, Chao Li, Pin Zhang, Zhang Xiong,A Query Expansion Algorithm based on Phrases Semantic Similarity, School of Computer Science and Technology, Beihang University, Beijing.

[13] C. de Loupy, P. Bellot, M. El-Beze and P.-F. Marteau, Query Expansion and Classification of Retrieved Documents