TREC 2009 at the University at Buffalo: Interactive Legal E-Discovery With Enron Emails

Jianqiang Wang
Department of Library and Information Studies
University at Buffalo, the State University of New York
Buffalo, NY 14260, U.S.A.
jw254@buffalo.edu

Ying Sun
Department of Library and Information Studies
University at Buffalo, the State University of New York
Buffalo, NY 14260, U.S.A.
sun3@buffalo.edu

Paul Thompson
General Dynamics Advanced Information Systems
Buffalo, NY 14225 U.S.A.
Paul.Thompson@GDIT.com

Abstract

For the TREC 2009, the team from University at Buffalo, the State University of New York participated in the Legal E-Discovery track, working on the interactive search task. We explored indexing and searching at both the record level and the document level with the Enron email collection. We studied the usefulness of fielded search and document presentation features such as clustering documents based on email threads. For query formulation for the selected search topic, we combined a precision-oriented Specific Query method that a recall-oriented Generic Query method. Future evaluation of the effectiveness of these query techniques is still needed.

1 Introduction of Legal Track's Interactive Task

For this year's TREC, the team from University at Buffalo (UB), the State University of New York continued to work on the interactive search task of the Legal E-Discovery track. Our main goals of working on the task include:

1. to further advance our understanding of legal e-discovery, in particular how document relevance is defined, perceived, and interpreted by lawyers, and
2. to study issues related to the design and the use of search systems for e-discovery with corporate emails.

The design of the interactive legal task is basically the same as last year [2]. Specifically, the task coordinators, topic authorities, document reviewers, and participating teams complete the task collaboratively in the following way:

- Track coordinators select topics

- Track coordinators select Topic Authorities (TAs) and assign topics;
- Participating teams work with TAs on topic clarification;
- Participating teams retrieve documents;
- Participating teams submit retrieval results;
- Document reviewers review sampled result documents;
- Track coordinators produce the first-pass evaluation results;
- Participating teams appeal relevance judgments that they disagree with or are unclear about;
- TAs adjudicate the disputed relevance judgments;
- Track coordinators produce the final evaluation results.

A new collection, namely the Enron email collection, is used for this year's interactive retrieval task. The collection was initially obtained from Aspen Systems and then processed by the track coordinating team. Preprocessing of the collection mainly includes extracting identifiable text and metadata and deduplicating messages. The resulting collection contains 569,034 unique messages and 267,131 attachments. Emails in this collection were originally created between 1998 and 2002. More details of the development of the document collection can be found in the track overview paper [1].

Seven search topics (i.e., request for production of responsive documents) were provided as part of the test collection for the interactive task. These topics were created for a hypothetical legal complaint of a securities fraud class action. In this case, the plaintiffs claim damages suffered from purchasing the company's common stock at artificially inflated stock prices and hence alleges a review of public documents, Securities and Exchange commission filings, analyst reports, news releases and media reports concerning the company. The seven search topics represent the plaintiffs' request for the defendants to produce responsive documents for the litigation. They seek documents about the company's engagement in prepay transactions, the company's engagement in transactions compliant with FAS 140, whether the company had met its financial forecasts, etc. For example, here is Topic 203, which is the topic that our team chose to work on:

- *All documents or communications that describe, discuss, refer to, report on, or relate to whether the Company had met, or could, would, or might meet its financial forecasts, models, projections, or plans at any time after January 1, 1999.*

One of the unique characteristics of the interactive legal task is the use of topic authorities (TAs). TAs play the role of senior litigators to define and clarify the responsiveness of documents regarding each request for production in legal e-discovery. Prior to the submission of its search results, each participating team can consult with the TA assigned to each topic about what makes a document relevant or not relevant using mutually agreed communication methods (except meeting face to face). In addition, TAs also provide relevance guidance for document reviewers and adjudicate relevance assessments appealed by participating teams.

Each participating team is required to retrieve responsive documents from the Enron email collection for at least one of the seven topics and submit its search results for official evaluation. However, each team is free to decide what search system to use, how much time it will spend on topic clarification with the TAs (as long as it is within the limit of 10 hours), and what search strategies it will use. Recall and precision of each submitted run will be estimated based relevance judgments of documents sampled from all submitted runs. The F measure that combines recall and precision will also be reported as part of the official evaluation results. Further information about the design of the 2009 TREC Interactive Legal Task can be found in the task guidelines [1].

The rest of the paper is organized as follows. In section 2, we present the design of our system for the interactive search task. Section 3 describes the techniques that we used to complete the

search task. Section 4 presents the evaluation of our run, including the appealing and adjudication result. Section 5 outlines several things that we have learned through our participation in the TREC Interactive Legal task and our plan of continuing the study of e-discovery.

2 System Design

For this year's interactive Legal task, we built a web-based search system that uses Indri as the back-end search engine. We chose Indri because it provides a variety of operators with which queries can be constructed and refined, in particular, query operators that support fielded search and quasi-Boolean retrieval. After examining some sample documents in the collection, we realized there were several things that needed to be considered for the design of the system. First, since it is quite common that many emails in the collection have attachments, we had to decide the unit of retireval, i.e., what makes up a *document*. An attachment can be treated as either part of its parent email or a separate document. These two treatments of email attachments correspond to the so-called "record-level" model in which each email and its attachments are treated as one document, and the "document-level" model in which emails and their attachments are regarded as independent documents. It was unclear which of these two models can lead to better retrieval effectiveness. Therefore, we decided to implement both models so that searchers could try them out. We hoped that evaluation of search results from these two models will provide us more insights of the usefulness of them.

The second thing we had to consider is how the structure of emails would be used. Emails are structured documents in that they contain metadata such as senders, receivers, carbon copied recipients, subject lines, time and dates, as well as email bodies. Each of these elements (i.e., fields) provides different information about the emails and we believe it is useful to provide searchers the capability of searching on any or all of them, i.e., fielded search. Prior to indexing, we identified these elements of emails and structured each email with field tags including FROM, TO, CC, DATE, SUBJECT, and TEXT. Accordingly, when formulating a query, a user can specify on which field(s) the search will be. Figure 1 shows the interface of our system that provides the fielded search capability. We call this interface "advanced search" interface. Meanwhile, we also created a "basic search" interface, with which the searcher does not need to specify which field(s) to search on. In that case, the query will be matched on all fields. For the "document model" in which each attachment is viewed as a document independent of its parent email, we simply populated the metadata fields (i.e., fields corresponding to the parent email's header section) to each of its attachments while treating the attachment content as the document body. This way, all attachments and their parents emails have the same structure. This pre-processing of the collection is unnecessary for the record-level model as the text of all the attachments are concatenated with the text body of their parent email to form a single document.

The presentation and display of search results of emails and their attachments is an interesting aspect of the system design. Email collections are different from news article collections but similar to blog collections in that the interrelationships among emails and attachments are often explicit. For example, emails in the same "thread" can usually be identified by the same subject line together with information of the sender, the receiver, and the time and date. For users, it could be useful to group emails from the same thread and present them as one cluster. Therefore, after receiving a ranked list from the backend Indri search engine, our system first groups the retrieved documents (emails and attachments) into clusters - each cluster represents one email thread - based on the highest rank in the original ranked list of documents in each cluster. As a result, each item showed in the initial search result page is actually a cluster of multiple documents. By clicking on the "show more documents from this thread" link the searcher will be able to see the list of documents in the selected cluster (see Figure 2 for a sample page with part of the search results.)

The document collection contains email attachments in their "native" formats (e.g., MS Word, Excel, etc.) and the converted text format. Ideally, we would want to have the proper Web browser plug-ins to directly display attachments in their native formats because that will make it

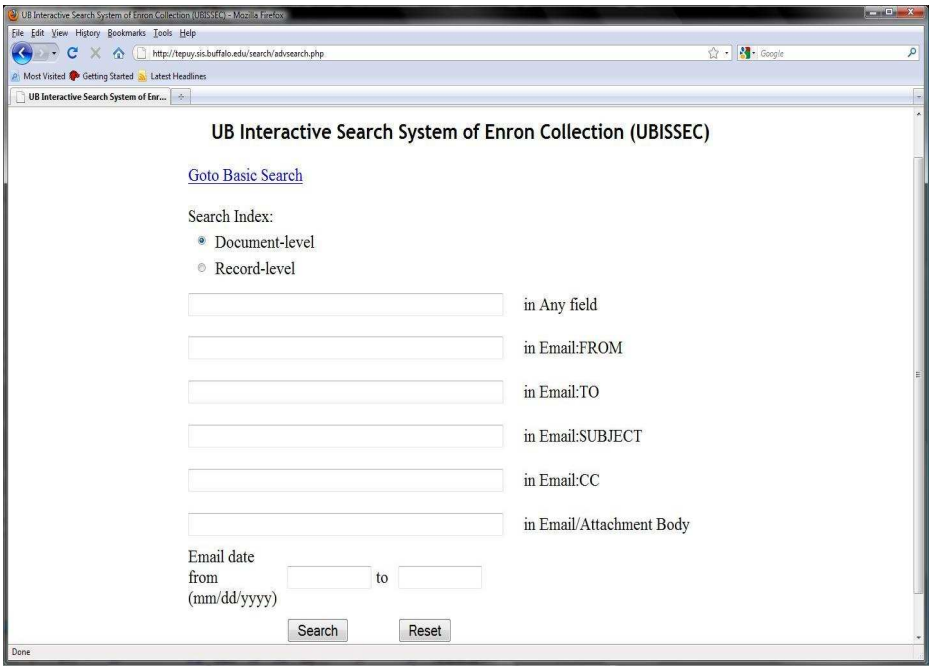

Figure 1: Advanced (field-based) search interface.

much easier for the searcher to read the attachments. Unfortunately, we did not implement this function due to time constraint. This system limitation did have some noticeable influence on both effective and efficient review of retrieved documents.

We also implemented a "bookbag" function in the search interface. A searcher can save the reviewed documents together with his relevance judgments in a bookbag, so that later he can check what queries he has tried and which documents he has reviewed. The searcher can also download the content of the bookbag to his local computer.

3 Search Techniques: Specific Queries and Generic Queries

After reviewing the task guidelines and the search topics and attending the Interactive task kickoff call, our team decided to work on Topic 203. We understood that a document should satisfy three conditions in order to be reviewed as relevant to the search request: (1) mentioning the company's financial forecasts, models, projections, or plans, (2) mentioning whether the company had met, could, would, or might meet them, and (3) a timeline after January 1, 1999. We were not sure, however, whether the time is about the document (email) creation date or it modifies the verb. Through the initial conference call with the TA of the topic, we realized that the time modifies the verb. Therefore, dates of emails in the collection is not reliable for deciding whether a document satisfies this condition. According to the guidelines given to the official relevance assessors, however, this time factor was not considered. Coincidentally, we did not do much on it either. Nevertheless, it is obvious that some kind of deeper linguistic analysis of email messages is required in order to resolve the time/date aspect for such search requests.

Some initial searches with the words contained in the topic statement returned very few relevant emails. Most of the retrieved emails talk about the company's financial models, forecasts, projections, and plans, but very few of them mention whether the company met the plan. We

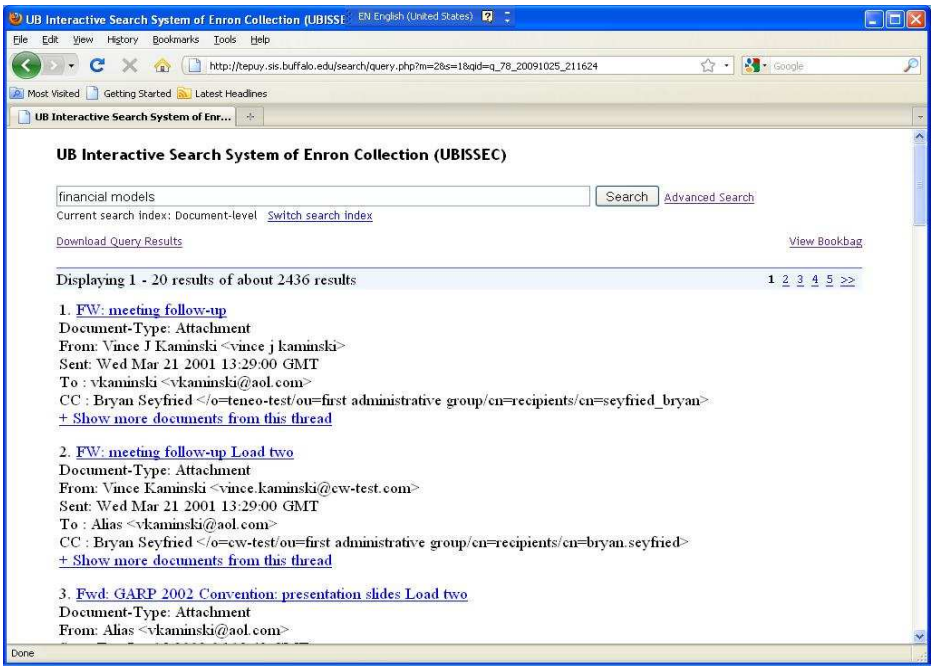

Figure 2: Display of search results.

realized that it would probably not help much if we limited our query words to those contained in the topic description. We consulted friends and other acquaintances who work in the business sector about colloquial expressions of companies meeting or not meeting their business plans. We also read related material such as the U.S. vs. Kenneth Lay et al Superseding Indictment to find out such useful expressions [1]. Base on what we learned, we formed a few strict queries and checked samples of retrieved documents to see if they were relevant. We called this query formulation technique *Specific Query* method because it focuses on retrieving documents based on key linguistic expressions or phrasal patterns. Indri's proximity operator #n is particular useful for formulating such queries. For example, query "#3(double avenue);" would match documents that contain word "double" and "avenue" in a window of 3 words (i.e., there is at most one word between the two words). Once we decided a string pattern is useful, we relaxed the restriction, i.e., the window size, so that more documents could be retrieved.

The specific query method could be effective in finding a few highly relevant documents, i.e., it is more precision-oriented. However, since high recall is of particular importance for legal e-discovery, we also applied a *Generic Query* method that could retrieve much more documents. The method puts many potentially useful words into a query, hoping the search engine will be able to rank them by factors such as term frequency and document frequency. It is more like queries used in a typical TREC ad hoc retrieval task. A main problem of the generic query method is that it tends to retrieve too many documents, hence one has to decide where to cut off the ranked list. In our experiment, we used a simple heuristic of sampling documents at some arbitrary rank region. If we felt most documents in that region were relevant, we moved forward to a lower rank region; otherwise, we moved back to a higher rank region. We repeated this process a few times until we felt we had reached the region of an optimal F measure, i.e., recall and precision are well balanced.

[1] http://news.findlaw.com/hdocs/docs/enron/usvlay70704ind.pdf

4 Experiment Results, Evaluation, and Adjudication

Query	Rtrvd	Rel	NonRel	Unsampled
#10(company strongest shape);	2	0	0	2
#10(company good shape);	5	0	0	5
#10(company great shape);	12	1	0	11
#10(company better shape);	6	0	0	6
#3(double revenue);	26	0	0	26
#1(quarter on quarter);	10	0	0	10
#or(#1(good year) #1(good quarter) #1(great year) #1(great quarter));	258	3	26	229
#10(revenue continue grow);	3	0	0	3
#10(consistent profitability);	50	0	6	44
#or(#1(high profitability) #1(higher profitability));	32	0	2	30
#2(hit number);	6	0	0	6
#1(strong growth);	355	7	30	318
#1(outstanding balance);	287	3	27	257
#band(enron financial quarter forecast);	1210	38	95	1077
#combine(raptors earnings);	4284	99	353	3832
#combine(enron finance forecast model projection plan analysis revenue profit increase decrease);	4547	37	437	4073

Table 1: **Queries and search result statistics.** Rtrvd: the number of documents retrieved by the query; Rel: the number of documents judged officially as relevant; NonRel: the number of documents officially judged as nonrelevant; Unsampled: the number of documents not officially sampled for relevance judgments.

For our official run, we constructed 13 specific queries and 3 generic queries. Each query returned a list of documents. All documents retrieved by used the specific queries were included in our official run. For ranked lists of documents returned by generic queries, we applied the cut-off heuristic described above. Finally, search results of the 16 queries were merged to form the official run of about 10,000 documents. According to the Interactive Legal task guidelines, document ranks were removed, i.e., the submitted run is a list of un-ranked documents.

Table 1 shows the queries we used for our official run and the final official evaluation of it. We can see for seven of the 13 specific queries, no document was officially reviewed. This is not surprising since those documents were merged with the much larger sets of documents retrieved with the generic queries. Therefore, they are much less likely to be sampled for official review. For the rest of the specific queries and all three generic queries, most of the retrieved documents were not sampled either. Again, that is understandable because sampling only took a small portion of the document pool for expert review. However, comparing the number relevant documents and that of nonrelevant documents, we can see that precision of these queries is poor.

After the first-pass assessments were released by the track coordinators, we selected 22 documents for adjudication by the topic authority. These documents were all initially judged as non-relevant by the document reviewers but we feel they are relevant. Due to the TA's time constraint, documents appealed for adjudication were limited to only those that were reviewed

in the first-pass assessments. It turned out that the initial relevance judgments of 16 of these 22 documents were reversed by the TA. After appealing and adjudication were completed, the track coordinators produced the final assessment results. Specially, the final effectiveness scores of our submitted run are recall of 0.592, precision of 0.111, and F of 0.186. Comparing them to the first-pass evaluation results of recall of 0.203, precision of 0.077, and F of 0.111, we see each measure is improved and in particular recall is almost tripled.

Along the way of completing the task, we also used other features of the system. Field-based search is helpful for quickly narrowing down the scope of retrieved documents. The bookbag is also useful for keeping the search history for later review purpose. Somehow we felt that the document-level model coupled with email threads was preferred over the record-level model, mainly because sometimes a retrieved record (i.e., the email plus its attachments) can be very lengthy, thus making it more difficult to quickly spot the relevant "nuggets." Therefore, all the queries in our submitted run were performed with the document model.

We did not keep exactly track of the amount of time we spent on the task. Based on our best estimation, it's somewhere between 60 and 100 man-hours for the search task alone, which does not include the time spent on designing, implementing, and debugging the system. Including the topic-specific conference call, we spent about 30 minutes with the TA.

5 Conclusion and Discussion

For this year's Interactive Legal E-Discovery task, our team built a prototype search system and tested the approach to combining specific queries and generic queries. Specific queries are more precision-oriented as they focus on the retrieval of documents that contain key phrases or other compounded linguistic expressions; generic queries are more recall-oriented because they usually contain many individual words but not more complicated linguistic units. Specific queries usually retrieves only a small number of documents while generic queries tend to return much more. Our experiment result supported this argument. However, it is still not validated that the specific queries we tried can achieve better precision mainly because for most of them either no document or only a very few documents were officially reviewed. It would be helpful in the cases like this to revise the design of the Interactive Legal task in the future so that teams are allowed to submit a limited number of documents that they definitely would like to be officially reviewed, in addition to the documents that will be pooled.

On the other hand, some of the specific queries we tried might be too specific. The large disparity between the number of documents retrieved by specific queries and the number of documents returned by generic queries in our experiment makes us believe there is a spectrum of queries with varied degrees of specificity between these two types of queries. It would be interesting to see how query specificity is related to effectiveness measures such as recall and precision. In addition, several unique features of email collections deserve more research attention, e.g., retrieval and browsing of emails at document level, record level, and even thread level. For the search topic we chose to work on, it is necessary to design techniques that can resolve the dates based on date-related terms (next quarter, last months, etc.) in email bodies and dates of emails. Also, it will be helpful to present users with email attachments in proper formats (MS Word, Excel, etc.) rather than in text format.

As electronic documents are becoming an important part of business records, narrowing the scope of documents to review for litigation purpose seems to be a necessity than an option. Undoubtedly, the kind of efforts made by the NIST and the TREC Legal track coordinators to bring together academic researchers and e-discovery practitioners have contributed to the continuing development of search technologies and techniques for legal e-discovery and the evaluation of the effectiveness and the usefulness of them. In addition to addressing the issues we identified above, we are planning to conduct a comprehensive study of comparing relevance judgments made by people with a law background and those without it. Our goal is to develop a relevance model that describes more accurately the different aspects of document relevance in e-discovery. Such a model can be used to guide the development of system features that can assist information searchers to

more quickly and precisely identify documents or sub-document elements for review.

6 Acknowledgement

The authors would like to thank the University of Massachusetts at Amherst and Carnegie Mellon University for making available the Indri search engine, UB's computer science student Ashok Narasimhan for implementing the search interface used in the study, and UB's former master of library science student Craig Preston for sharing his legal knowledge. This project is partly supported by the UB Baldy Center for Law and Social Policy through an annual research grant.

References

[1] Jason R. Baron, Bruce Hedin, Douglas W. Oard, and Stephen Tomlinson. Interactive task guidelines: 2009 TREC legal track. 2009. Available at: http://trec–legal.umiacs.umd.edu/LT09_Complaint_J_final.pdf.

[2] Douglas W. Oard, Bruce Hedin, Stephen Tomlinson, and Jason R. Baron. Overview of the TREC 2008 legal track. In *The Sevententh Text REtrieval Conference*. National Institutes of Standards and Technology, November 2008. Available at: http://trec.nist.gov.