

# **UB at TREC 2010 Legal Interactive**

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

## **Abstract**

For the TREC 2010, the State University of New York at Buffalo participated in the Legal E-Discovery task, working on the interactive search task. We selected to explore RPD task 303. Our focus was on how to approach the problem with the assumption that business communication often wants to maintain secrecy or plausible deniability. Accordingly, it is not in the spirit of the problem to approach formulating queries by limiting ourselves to the mere text of the Complaint and RPD's. We have to envision the actual business context and the actual business practices to determine truly effective queries in the context of litigation. A simple interactive system based on Indri search engine was used to test the queries and examine the results. Post-experiment analysis is underway in alignment with the official evaluation.

## **1 Introduction**

In TREC 2010 Legal Track, the team at the State University of New York at Buffalo (UB) team continued to participate in the Interactive Task. With limited resources, we selected to work on in only one topic request: topic 303 (lobbying).

The main purpose of the Interactive Task is to model more completely and accurately the task of reviewing documents for responsiveness to a request for production in civil litigation [2]. The design of the interactive legal task is modeling the real-world conditions in which e-discovery is conducted collaboratively by searchers (the participating teams) and the senior litigators (the topic authorities). The senior litigators (TAs) play a very important role in the e-discovery process by defining the scope of responsiveness. However, their available time and effort in the e-discovery process is generally very limited. Well prepared searchers can communicate with the senior litigators more efficiently. Our interest is to explore the e-discovery process from the searchers' point of view.

## **2 Collection**

For this year's interactive retrieval task, a new version of the Enron email collection is used. The new version is provided by the Electronic Discovery Reference Model (EDRM). The EDRM Enron Email Data Set v2 consist of Enron e-mail messages and attachments in two sets of downloadable compressed files: XML and PST. The XML collection was further cleaned up by the track coordinating team. The resulting collection contains 685,592 unique email messages

and attachments. The emails were originally created from about 150 users, mostly senior management of Enron between 1998 and 2002, organized into folders. This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission (FERC) during its investigation.

A Mock complaint with three responsiveness topics and one privilege topic was developed for this year's legal track. The complaint document sets forth the legal and factual basis for the hypothetical lawsuit that motivates the discovery requests. We chose to participate in one document responsiveness topic: topic 303.

*303. All documents or communications that describe, discuss, refer to, report on, or relate to activities, plans or efforts (whether past, present or future) aimed, intended or directed at lobbying public or other officials regarding any actual, pending, anticipated, possible or potential legislation, including but not limited to, activities aimed, intended or directed at influencing or affecting any actual, pending, anticipated, possible or potential rule, regulation, standard, policy, law or amendment thereto.*

### 3 Searching Systems

We built a very simple web-based interactive search system. It uses Indri as the back-end search engine. We chose Indri because it provides rich query operators with which queries can be refined, for example, field-based search and quasi-Boolean queries. The interactive system is designed purely for the purpose of trying queries and reviewing documents. A simple query box is used to submit Indri queries (Figure 1).

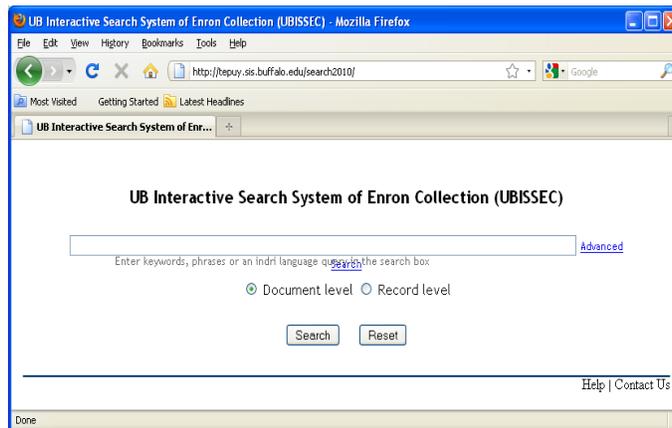
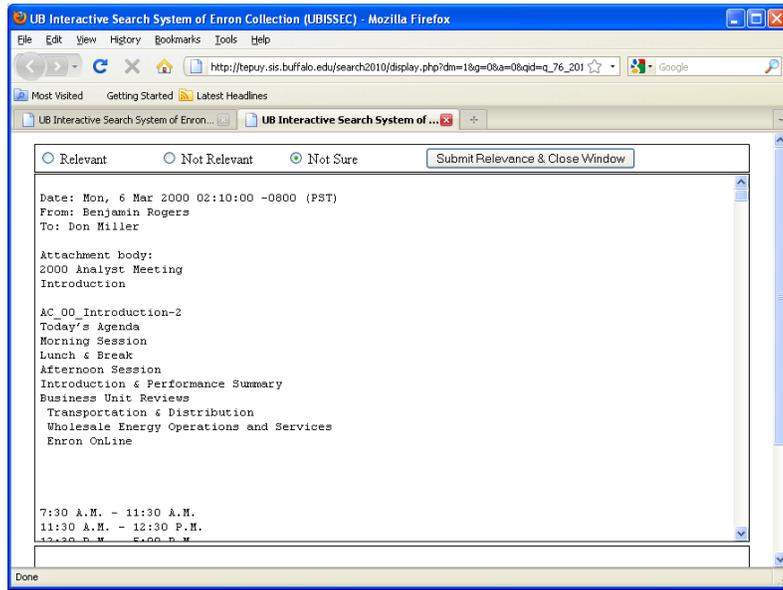


Figure 1. Interactive System Query Interface

To facilitate the document reviewing process, a "bookbag" function is implemented in the interactive system. Along the way of trying queries and reviewing documents, a searcher can save the reviewed documents together with her/his relevance judgments in a "bookbag"(Figure 2), so that later she/he can check the queries and the documents she/he has saved. A searcher can also download the content of the "bookbag" to a local computer. More details about the system can be found in [3].

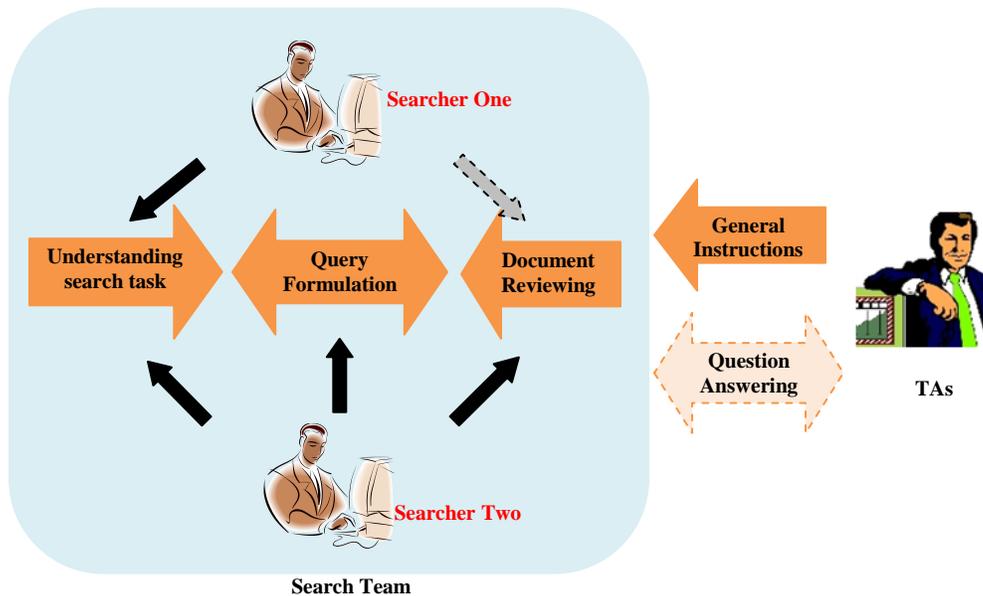


**Figure 2 Bookbag Function**

The batch mode Indri engine was used finally to retrieve documents given queries generated through the interactive process.

#### 4 Task Execution

We have two searchers initially in our search team. Our task execution plan is illustration in Figure 3.

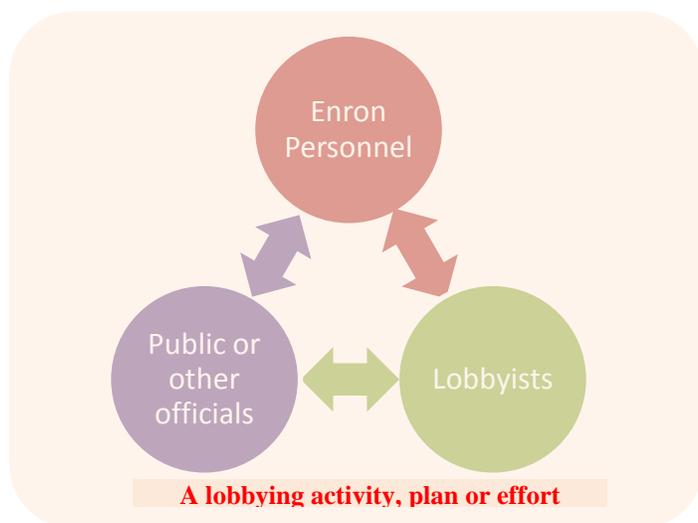


**Figure 3 Task Execution Process**

Searcher one is a scholar with legal training background. However, he is not an expert in either e-discovery or the case related area. This searcher was expected to provide initial suggestions about how to approach to the topic at the very beginning of the task and to provide suggestion throughout the whole task process by reviewing retrieved documents. However, after providing the initial suggestion, due to unexpected reason, the first searcher could not continue his work on the task. The second searcher is an expert searcher with LIS training background, who is the major searcher who worked together with the first searcher to understand the task, then generated queries and reviewed retrieved documents. The whole search team got general instructions from the TA. We also planned to take full advantage of the 10-hour permitted TA time for specific questions and clarifications. However we could only conduct very limited communication with TAs due to time and human power constrains. In the illustration in Figure 3, the un-fulfilled parts are illustrated with dash boarder.

We believe that, given that business communication often wants to maintain secrecy or plausible deniability, it is not in the spirit of the task to approach formulating keywords by limiting ourselves to the mere text of the Complaint and the RPD's. We have to envision the actual business context and the actual business practices to determine truly effective keywords in the context of litigation.

By analyzing the request topic, the search team developed a topic model for the searching task (Figure 4). Three possible parties are identified to be involved in any activities, plans or efforts (whether past, present or future) aimed, intended or directed at lobbying.



**Figure 4 Parties in a Lobbying activity, plan or effort**

We decided to search by identifying personals in these three parties. External web resources were consulted. Next Table summarized persons we identified.

<p><b>Enron Personnel:</b>  James Steffes Vice President Government Affairs  Jeff Dasovich Director for State Government Affairs.  Richard Shapiro Vice President Regulatory Affairs  Steven Kean Vice President Vice President &amp; Chief of Staff  Susan Mara  Government affairs (GAs)</p>
<p><b>Lobbyists:</b>  Ralph Reed, Bev Hansen, Hedy govenar, Scott Govenar, Mike Day, Robin gibbs,  Carl Ecklund, Gary Fergus, Dan Watkiss, Theo Pahos</p>
<p><b>Officials:</b>  <b>FERC:</b>  Curt Hebert (Chairman)  Pat Wood (Chairman)  Joseph Kelliher (Chairman)  Dan Larcamp (Director, Office of Markets, Tariffs and Rates (“OMTR” )  <b>U.S. Department of Energy (DOE)</b>  Spencer Abraham (Secretary)  Bill Richardson  Jimmy Glotfelty, Senior Electric Policy Advisor to Energy Secretary Spencer  Abraham DOE;  Frank Blake, Deputy Secretary DOE;  <b>CA</b>  Gray Davis (governor)  <b>Texas</b>  Rick Perry  Public Utility Commission of Texas – PUCT  <b>The Nuclear Regulatory Commission (NRC)</b>    Andrew Lundquist (Senate Energy Committee/White House Task Force)  Bill Condit (House Energy and Mineral Resources subcommittee)  Stacey Gerard (Office of Pipeline Safety)</p>

**Table 1. Names used in Queries**

For each identified name or organization, we tried various possible ways it may be referred in the email collection. Only the promising ones were kept in the final query sets. When we formulated the Indri queries, NOT operator was used to exclude irrelevant documents that may be retrieved by promising query words.

We also tried some queries based on lobbying activities, for example, CA energy crisis, China WTO, etc.

51,545 documents were retrieved and submitted based on the manually generated queries using Indri search engine.

## 5 Results

The results are summarized in the table below. The second column is the preliminary (pre-adjudication) estimates of the recall, precision, and F1 achieved by UB's submission on topic 303. On this topic, there were submissions in total; the average recall, precision, and F1 of the six submissions are listed in the third column.

	UB Initial	Initial Average	UB Final
Recall	.54	.34	.72
Precision	.41	.57	.30
F1	.34	.38	.42

**Table 2. Results**

802 messages were submitted during the appeal process. In the adjudication set, 304 (37.9%) saw a change in assessment as a result of the adjudication process. This year the TREC organizer also included in the adjudication set a selection of non-appealed messages. Of the 187 non-appealed messages in the adjudication set, 42 (22.5%) saw a change in assessment as a result of the adjudication process. Of the 989 total messages in the adjudication set, 346 (35.0%) saw a change in assessment as a result of the adjudication process. UB's final scores are listed in the last column.

## 6 Discussions

In this exercise, we were simulating the very first step action in our proposed e-discovery working model. With limited efforts (less than 100 hours work in total and little interaction with TAs), our exercise achieved above average performance. What we learned from this exercise is that the importance of external resources about the topic. These recourses were used in two ways: 1. as the knowledge base to help the searchers to understand the topic and think about the topic in various ways; 2. as the sources of query words. We are planning to further study the importance of domain knowledge of the topic in the e-discovery process to find out how it is used with the final goal of automatic integration of such external resources into the search process.

## References

- [1] TREC 2010 Legal Track Interactive task guidelines. Available at: [http://trec-legal.umiaccs.umd.edu/itg10\\_final.pdf](http://trec-legal.umiaccs.umd.edu/itg10_final.pdf)
- [2] Douglas W. Oard, Bruce Hedin, Stephen Tomlinson, and Jason R. Baron. Overview of the TREC 2008 legal track. In *The Seventeenth Text REtrieval Conference*. National Institutes of Standards and Technology, November 2008. Available at: <http://trec.nist.gov>.

[3] Jianqiang Wang, Ying Sun. TREC 2009 at the University at Buffalo: Interactive Legal E-Discovery With Enron Emails. In *The Eighteenth Text REtrieval Conference*. National Institutes of Standards and Technology, November 2009. Available at: <http://trec.nist.gov>.