Indian Statistical Institute, Kolkata at TREC 2010 : Legal Interactive

Kripabandhu Ghosh[1], Swapan Kumar Parui[1], Prasenjit Majumder[2],
Ayan Bandyopadhyay[1] and S. John J. Raja Singh[3]

[1] Indian Statistical Institute, Kolkata, West Bengal, India

[2]Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat, India

[3]Indian Institute Technology, Kharagpur, West Bengal, India

Abstract

Indian Statistical Institute, Kolkata participated in TREC for the first time this year. We participated in TREC Legal Interactive task in two topics namely, Topic 301 and Topic 302. We reduced the size of the corpus by Boolean retrieval using Lemur 4.11[1] and followed it by a clustering technique. We chose members from each cluster (which we called *seeds*) for relevance judgement by the TA and assumed all other members of the cluster whose seeds are assessed as relevant to be relevant.

1 Introduction

The motivation for our participation in TREC Legal track was to test the conventional tools of Information Retrieval on legal data. We chose the Interactive task [3] because of its uniqueness as in contrast with any other Ad Hoc task it allows the participants to interact with the reviewer as a process to understand the notion of relevance/responsiveness [1]. The data set this year was the EDRM Enron v2 dataset which consisted of Enron emails and their native attachments separately provided. There were two formats of the data on offer viz. XML and PST. Later on deduplicated text-only version was also available which we chose for our experiment. The data was available at http://durum0.uwaterloo.ca/trec/legal10/. The emails were of 596MB (compressed) and the native attachments were of 6GB (compresssed). We used Indri search engine of Lemur 4.11 toolkit for Boolean retrieval and Terrier 3.0[2] for ranking the retrieved set using DFR [2]-BM25 [4] model. In section 2, we describe our approach. In section 3, we present the results and we conclude in section 4.

2 Our Approach

We attempted to apply DFR-BM25 ranking model on the TREC legal corpus. We chose Terrier 3.0 as this toolkit has most of the IR methods implemented within. But as we received the TREC legal data set we realised that it would be difficult to manage such a large volume of data. So, we decided to reduce the corpus size by Boolean retrieval. We chose Lemur 4.11 as it supports various useful Boolean query operators which would suit our purpose. On the set obtained from Boolean retrieval we decided to apply ranked retrieval techniques. We decided to index only the original emails. The attachments were not indexed. We

[1]http://www.lemurproject.org/
[2]http://terrier.org/

decided to add the attachments of the relevant parent emails in the final result because we assumed that the attachments of a relevant parent mail are likely to be relevant as well. The use of Boolean retrieval has the disadvantage that it will limit further search to the documents retrieved at this stage and have an adverse effect on our recall performance. But it would scale down the huge corpus size considerably (see Table 1) and enable us to perform our experiments on a smaller set which would reduce processing time.

2.1 Topic 301:

Topic 301 was as follows:

```
All documents or communications that describe, discuss, refer to, report on, or relate to
onshore or offshore oil and gas drilling or extraction activities, whether past, present or
future, actual, anticipated, possible or potential, including, but not limited to, all
business and other plans relating thereto, all anticipated revenues therefrom, and all risk
calculations or risk management analyses in connection therewith.
```

Based on the topic, we formed a Indri [3] query which combines the result of the following eight queries:

1. #or(#band(oil gas onshore drilling anticipated revenue risk explosion fire) #band(oil gas offshore drilling anticipated revenue risk explosion fire) #band(oil gas onshore extraction anticipated revenue risk explosion fire) #band(oil gas offshore extraction anticipated revenue risk explosion fire))
2. #or (#band(#1 (oil and gas) drilling onshore rig anticipated revenue risk) #band(#1 (oil and gas) extraction onshore rig anticipated revenue rig risk) #band(#1 (oil and gas) drilling offshore rig anticipated revenue risk) #band(#1 (oil and gas) extraction offshore rig anticipated revenue risk))
3. #or (#band(oil gas drilling onshore) #band(oil gas drilling offshore) #band(oil gas extraction onshore) #band(oil gas extraction offshore))
4. #filreq (#band(oil and gas) #combine(onshore offshore drilling extraction revenue risk))
5. #or (#band(#1 (oil drilling) onshore) #band(#1 (oil extraction) onshore) #band(#1 (oil drilling) offshore) #band(#1 (oil extraction) offshore) #band(#1 (gas drilling) onshore) #band(#1 (gas extraction) onshore) #band(#1 (gas drilling) offshore) #band(#1 (gas extraction) offshore))
6. #or (#band (#30(onshore oil drilling) revenue) #band(#30 (onshore oil extraction) revenue) #band (#30 (onshore gas drilling) revenue) #band (#30 (onshore gas extraction) revenue) #band (#30(offshore oil drilling) revenue) #band (#30 (offshore oil extraction) revenue) #band (#30 (offshore gas drilling) revenue) #band (#30 (offshore gas extraction) revenue))
7. #or (#30(onshore oil drilling) #30 (onshore oil extraction) #30 (onshore gas drilling) #30 (onshore gas extraction) #30(offshore oil drilling) #30 (offshore oil extraction) #30 (offshore gas drilling)#30 (offshore gas extraction))
8. #band (#1 (#syn(oil gas) #syn(drilling extraction)) #syn(onshore offshore))

This query returned a set of 2896 documents as compared to 685592 (a reduction to 0.42%) in the given collection of emails. The comparison is given in table 1.

Then we had to select a sample from the retrieved set for TA judgement. Instead of picking random samples, we decided to rank the retrieved set based on a Terrier query and select the top ranked documents for TA assessment. So we formed a the following Terrier query :

⟨ *title*⟩ Topic: Oil or Gas Drilling or Extraction

⟨ *desc*⟩ Description:

[3] http://www.lemurproject.org/lemur/IndriQueryLanguage.php

Collection	No. of documents	Collection size
Original corpus	685592	3.8GB
Collection after Boolean retrieval	2896	227.8MB

Table 1: Collection statistics for Topic 301

Document describes oil and gas onshore drilling or extraction activities, business plans, revenues and risk management

⟨ *narr*⟩ Narrative:

To be relevant, a document should describe onshore or offshore oil and gas drilling or extraction activities all business and other plans related, all anticipated revenues, and all risk calculations or risk management analyses

We applied DFR-BM25 model of Terrier 3.0 using the above query on the aforesaid set obtained by Boolean retrieval of Lemur. We chose the top 10 documents of the resulting ranked-list for Topic Authority assessment. In our first discussion with TA for Topic 301 (Mira Edelman), we offered these 10 documents for her opinion. We further requested for a clearer notion of relevance for the topic. Only three of the documents were judged relevant:

- Enron oil/gas drilling/extraction activities and revenues earned therefrom
- Enron oil/gas evaluation/audit report
- Enron oil/gas transaction/ agreement/ transportation would be relevant.

***Gas agreement* documents:** According to the TA, any document about a oil/gas agreement between Enron and another company would be relevant. One of the judged documents about Gas agreement as cited out by the TA was of the form:

```
ENFOLIO EXCESS GAS PURCHASE AGREEMENT
(RESERVES COMITTED/INDEX PRICING)
Enron North America Corp. ......
```

To capture documents with a similar structure we formed a query :

q301-1 : #5(#1(gas purchase agreement) reserves committed index pricing #syn(ena enron))

We used our *clustering algorithm* (to be discussed in section 2.3) to form clusters. We sent one member document from each, which we call a *seed*, for TA assessment. Five documents were judged relevant and their clusters were added to our set of relevant documents. This was a far too focussed query. To capture other *agreement* documents we formed another Boolean query (q301-2) :

#uw10(#syn(oil gas) #syn(ena enron) #1(purchase agreement))

As we started interacting with the TA, we realised that instead of making a two stage retrieval (i.e. Boolean followed by ranked) we could get better results by Boolean retrieval alone if we manage to choose appropriate keywords. On the Boolean retrieved set, we decided to apply clustering instead (see section 2.3 for details). This technique seemed to work well as judging by TA feedback we seemed to get more relevant documents. In the wake of this, we decided to use TA advice as feedback and did not make use of any relevance feedback technique. So, in the remainder of our experiments we stuck to Boolean followed by clustering approach.

Collection	No. of documents	Collection size
Original corpus	685592	3.8GB
Collection after Boolean retrieval	2715	225.7MB

Table 2: Another collection statistics for Topic 301

***Risk management* documents:** To retrieve documents regarding Enron risk management issues we formed the following two Boolean Lemur queries:

q301-3 : #band(#1(enterprise risk management) #syn(ena enron) risk asset audit)

q301-4 : #band(#syn(oil gas) #syn(enron ena) risk #(operational risk) asset audit)

The key terms "enterprise risk management", "operational risk" etc were suggested by both TA feedback and Rocchio Relevant feedback technique in Terrier 3.0.

***Enron audited report* documents:** To obtain the documents related to Enron's audited report on net income related to oil/gas activities we formed the query:

q301-5 : #band(#1(oil and gas) #syn(ena enron) #1(financial statement) asset liability equity income expense audit #1(consolidated net income))

***Oil and gas transportation* documents:** As suggested by the TA, we looked for the documents about Enron oil and gas transportation activities by the following query :

q301-6 : #band(#1(#syn(oil gas) #syn(drill extraction)) #syn(ena enron) #5(#syn(oil gas) transport))

The last three queries were not high yielding. So we decided to make a large set of all drilling-extraction documents. We formed the Lemur query :

q301-7 : #band(#1(oil and gas) #syn(ena enron) #syn(drilling extraction))

This again shrunk the original corpus as given in table 2.

As before, we made a ranked retrieval with DFR-BM25 using the Terrier query:
⟨ *title*⟩ Topic: Enron Oil or Gas Drilling or Extraction
⟨ *desc*⟩ Description:
Document describes Enron oil and gas onshore drilling or extraction activities
⟨ *narr*⟩ Narrative:
To be relevant, a document should describe Enron onshore or offshore oil and gas drilling or extraction activities

We produced manually picked documents from the set thus obtained for TA assessment. The following subcollections evolved from the above set of retrieved documents after TA feedback:

1. Enron weekly summary - news about Enron's business news
2. Enron Btu weekly summary (news about Enron's internal affairs)
3. Enron SIC codes - This category encouraged fresh search from the whole collection by the Boolean query:

   q301-8 : #band(#1(oil and gas) #syn(ena enron) #1(sic code))

4. Enron stock
5. Enron Austin

***Pad Gas* documents :** A document retrieved by q301-6 (Oil and gas transportation) talked about Pad Gas which led to the query :

q301-9 : #band(#syn(oil gas) #syn(ena enron) #1(pad gas))

Further analyses led to the following types:

***Texas business plan* documents:** q301-10 : #band(#syn(oil gas) #syn(ena enron) #1(texas gas) #1(business plan))

***Competitive analysis* documents:** Another document retrieved by q301-6 (Oil and gas transportation) led to
q301-11 : #band(#syn(oil gas) #syn(ena enron) #1(competitive analysis))

***California Energy Commission* documents:** Another document retrieved by q301-6 (Oil and gas transportation) prompted us to form
q301-11 : #band(#syn(oil gas) #syn(ena enron) #syn(drilling extraction) revenue #1(california energy commission))

***Global Contracts* documents:** We intended to look for the documents about the Global oil/gas contracts of Enron. So we came up with the query:
q301-12 : #band(#syn(oil gas) #syn(ena enron) #1(global contract) #syn(purchase sale transport) financial)

***Confidential Enron* documents:** In one of the later calls, TA suggested that the documents containing Enron's confidential news or reports about oil/gas are to be relevant. This led us to form the following query:
q301-13 : #band(#syn(oil gas) #syn(ena enron) #20(confidential propriety enron internal))

***Oil purchase* documents:** Finally we looked to grasp the probable left out documents by a query:
q301-14 : #band(enron #1(oil purchase) agreement)

2.2 Topic 302:

Topic 302 was as follows:

```
All documents or communications that describe, discuss, refer to, report on, or relate to
actual, anticipated, possible or potential responses to oil and gas spills, blowouts or
releases, or pipeline eruptions, whether past, present or future, including, but not limited
to, any assessment, evaluation, remediation or repair activities, contingency plans and/or
environmental disaster, recovery or clean-up efforts.
```

We started our experiments with Topic 301 where we noted that query formation merely from the keywords of the topic can be misleading in reaching the relevant documents. So, in case of Topic 302 we didn't form different combinations of Indri queries and instead tried to get hold of a few documents likely to be useful for our first interaction with the TA. So, we started with the Boolean-ranked strategy. We formed an Indri query :

Collection	No. of documents	Collection size
Original corpus	685592	3.8GB
Collection after Boolean retrieval	1350	17.4MB

Table 3: Collection statistics for Topic 302

#band(#syn(enron ena) #syn(oil gas) #syn(spill blowout release eruption))

The Boolean query yielded 1350 documents. The number is tabulated in table 3. Thus the corpus was reduced to 0.197%.

We formed a ranked-list using Terrier 3.0 DFR-BM25 using the query:

⟨ *title*⟩ Topic: Oil and Gas Spills

⟨ *desc*⟩ Description:

Document describes oil and gas spills, blowouts,releases,pipeline eruptions and assessments,repair or remediation

⟨ *narr*⟩ Narrative:

To be relevant, a document should describe oil and gas spills, blowouts,releases,pipeline eruptions assessment, evaluation,remediation or repair activities,contingency plans and/or environmental disaster,recovery or clean-up efforts

Top 10 queries were presented to TA of Topic 302 (John Curran) on our first call. This interaction revealed a gross misinterpretation of the notion of relevance on our part as none of the 10 retrieved documents were deemed responsive!

At this stage, we had decided on using our clustering algorithm and so, we will be using this strategy in the remaining part of Topic 302.

***Clean up* documents:** TA opined that any document about "clean up effort" of oil/gas related to Enron would be responsive. These are the documents which narrate that some gas/oil spill incident caused by Enron has taken place and the consequent clean up efforts have been initiated. So we reformulated our query as:

q302-1 : #band(enron #syn(oil gas) #syn(spill release) #uw10(#1(clean up) spill))

As clean-up measures would involve the application of tools like skimmers, booms and chemical dispersants, we formed a query as follows:

q302-2 : #band(#syn(enron ena) #syn(oil gas) #uw100(spill #syn(skimmer boom dispersant))

***River spill* documents:** To capture documents about oil/gas spills in river areas we formed query:

q302-3 : #band(enron #syn(oil gas) spill river)

***Transredes spill* documents:** To target documents about Transredes spill we came up with

q302-4 : #band(enron #syn(oil gas) spill transredes)

***Litigation Memorandum* documents:** For the documents about litigations about oil/gas spill against Enron our query was:

q302-5 :#band(enron #syn(oil gas) spill litigation memorandum)

***Action Plan* documents:** During one interation, the TA also suggested that the documents about action plan of Enron about the prevention/remedial efforts in case of oil/gas spills would be responsive. This prompted us to form the following query:

q302-6 :#band(enron #syn(oil gas) spill #1(action plan))

***Spill Environmental* documents:** To retrieve the documents about environmantal hazards caused by oil/gas spill by Enron and the legal actions taken in this issue we formed the query:

q302-7 :#band(enron oil spill #10(environmental #syn(law matter)))

Topic number	Time expended with TA	No of documents retrieved as relevant(tentative)
301	2hours 38 mins	693
302	1 hour 37 mins	109

Table 4: Time with TA and size of retrieved set

query no	no of seeds	no of docs in relevant clusters
q301-1	4	79
q301-2	5	197
q301-3	4	8
q301-4	7	12
q301-5	2	12
q301-6	3	9
q301-8	4	21
q301-9	3	16
q301-10	1	3
q301-11	7	67
q301-12	1	20
q301-13	12	122
q301-14	2	62
q302-1	7	45
q302-2	2	3
q302-3	6	30
q302-4	2	6
q302-5	1	6
q302-6	1	4
q302-7	1	19

Table 5: Seeds and Clusters

2.3 Clustering Algorithm

On the basis of emperical studies we chose 0.3 as the *threshold* value. Initially, we provided most of the retrieved documents for judgement. But, gradually we observed that there exist many clusters of very similar documents and the relevance of all the documents in such a cluster can be decided by judging a few

Run status	Est. Recall	Est. Precision	Est. F_1
Pre-appeal	0.017	0.643	0.033
Post-appeal	0.027	0.867	0.052
% Improvement on appeal	58.82	34.84	57.58

Table 6: Results : Topic 301

members belonging to it. We believed that a document similar to a relevant document is likely to be relevant proportionally with the degree of similarity. So we formed clusters based on cosine similarity and tested out our assumption through relevance judgements. Positive results encouraged us to go on with it. The formal algorithm is as follows:

Let $G(V, E)$ be an undirected graph, where V (the set of vertices) is the set of all documents in a given collection C. There is an edge $e \in E$ between vertices $v_1(d_1)$, $v_1(d_2) \in V$, d_1, d_2 being documents of C, if the normalised *cosine similarity* between d_1 and d_2 is greater than *threshold* (In our experiments, *threshold* is chosen as 0.3). Next, the *connected components* of G are found out. These components are our clusters. This is basically a *single-linkage clustering* which we thought would be appropriate for our experiment.

A cluster containing one or more judged relevant document(s) is considered as a "relevant cluster". In other words, each document of the cluster is assumed to be relevant. For a cluster not containing a judged relevant document, we send a few arbitrarily chosen documents as the representatives (or seeds) of the cluster for TA judgement. Such a cluster will be deemed relevant or nonrelevant according as its seeds are relevant or not.

3 Results

We believe that the chances of achieving better understanding of the notion of relevance of a legal topic is directly proportional to the number of hours expended with the Topic Authority. Our team spent more time with TA of Topic 301 and managed to retrieve more useful documents. Table 4 illustrates this notion.

Table 5 shows the results of using the clustering algorithm and generation of more tentative relevant documents starting from a relatively small number of seeds. These results depict the performance of high yielding topics like q301-2, q301-13, q301-1. For the low yielding topics, clustering technique was of little use.

The results of Topic 301 are shown in Table 6. After the first-pass sampling of Topic 301, 140 of the documents submitted by our team were selected for assessment. There were 42 documents on which we differed with the assessors. We appealed against 18 of them and 15 of them went in our favour. The sampling and adjudication details is presented in tabular form in table 7.

Docs submitted	Sampled for assessment	Disagreements	Appealed against	Appeals won	% Appeals won
1394	140	42	18	15	83.33

Table 7: Sampling and Adjudication : Topic 301

The results of Topic 302 are shown in Table 8. After first-pass sampling, 267 of the 274 documents submitted by us were chosen. There were 171 documents which we thought were Responsive and the assessors had deemed them as Non-Responsive. We appealed against 37 of them and got 22 overturned (see table 9).

Run status	Est. Recall	Est. Precision	Est. F_1
Pre-appeal	0.054	0.476	0.097
Pre-appeal	0.090	0.693	0.160
% Improvement on appeal	66.67	45.59	64.95

Table 8: Results : Topic 302

Docs submitted	Sampled for assessment	Disagreements	Appealed against	Appeals won	% Appeals won
274	267	171	37	22	59.46

Table 9: Sampling and Adjudication : Topic 302

4 Conclusion

This participation was a great learning experience for our team. Resource constraints and time constraints were major challenges on our part. We believe that we could have made much better use of TA assessment as we managed to interact with TAs for a period of one month. We came up with a clustering technique which was applied on the output of Boolean search to help us maximize the benefit of these interactions. We decided to submit clusters of relevant documents, about which we had high degree of confidence from TA feedback. Our high success in appeal (83.33% for Topic 301 and 59.46% for Topic 302) shows that we managed to capture the notion of relevance to a considerable degree. In a nutshell, we prepared a precision based system aimed at capturing the relevant documents. Our high precision values attest to the fact that we have succeeded in our approach to a great extent. But, it seems that this has come at the expense of recall. This reason behind this is the fact that we worked with a very small sub-collection of the given corpus.

We feel that reducing expert dependence (here TA) would speed up the retrieval process. So, it is worthwhile to look to automate the process of query formation by query expansion tools that make use of legal knowledge. Doing away with a human expert competely may reduce system reliability and completeness. So, we may look to achieve more guidence in lesser number of interactions. We used a single linkage clustering which are not without its weaknesses. So, we hope to apply a better algorithm. Also, we hope to make a comparison of Boolean and ranked retrieval techniques. Finally, we would like to strike a balance between precision and recall values. This may be achieved if we work with the whole corpus instead of shrinking it.

5 Acknowledgements

We are honoured to acknowledge the kind contributions of Dr. Mandar Mitra, Indian Statistical Institute, Kolkata, India for his valuable advice on IR methodologies and Dr. Shreya Matilal, Rajiv Gandhi School of Intellectual Property Law, Indian Institute of Technology, Kharagpur, India for his vision in the legal domain. We also heartily thank the Topic Authorities - Mira Edelman (Topic 301) and John Curran (Topic 302) who were extremely cooperative with our team.

References

[1] Trec2010 legal track interactive task guidelines. Available at : http://trec-legal.umiacs.umd.edu/itg10_final.pdf, 2010.

[2] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, 2002.

[3] B. Hedin, S. Tomlinson, J.R. Baron, and D.W. Oard. Overview of the trec 2009 legal track. Available at : http://trec.nist.gov/pubs/trec18/t18_proceedings.html, 2009.

[4] Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.