Cluster-based Relevance Feedback : Legal Track 2011

Kripabandhu Ghosh[1], Prasenjit Majumder[2] and Swapan Kumar Parui[1]

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

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

Abstract

This is our second participation in the TREC Legal Track. The TREC Legal Track 2011 featured only the Learning Task. We participated in Topics 401 and 403. We used Lemur 4.11[1] for Boolean retrieval and followed it with a clustering technique, where 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. Based on the relevance information from seeds and their clusters, we applied Rocchio relevance feedback technique implemented in Terrier 3.0[2]. Then, we used the feedback terms for the expansion of both the text queries and the Boolean queries. Finally, we used Z-fusion[4], a data fusion technique, on two of our runs.

1 Introduction

This year, TREC Legal Track comprised of a single task - Learning Task, which was similar to the 2010 Learning Task except that this year the participants would start with "zero knowledge" (no training documents were available). Three topics - nos 401, 402 and 403, were given for the task. For each topic, the assigned Topic Authorities (TAs) provided "coding guidelines" and conducted a "kick-off" call, which was the only opportunity for the teams to interact with the TAs. Our team participated only in topics 401 and 403. The participants had to submit interim runs on the basis of the number of documents submitted for assessment, before the final submission. They could ask for relevance assessments of maximum 1000 documents per topic, at most 100 documents each time. After the final submission, the relevance judgements of all the documents requested for determination to the TA by all the teams were made available. The participants could use these determinations and make mop-up submissions. The data set this year was the same as that used in the last year - 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 (compressed). The collection contained 685592 documents. We used Indri search engine of Lemur 4.11 toolkit for Boolean retrieval and Terrier 3.0 for Rocchio relevance feedback using DFR [2]-BM25 [3] model.

We describe our approach in section 2, present our results in section 3 and conclude in section 4.

2 Our Approach

In our last year's participation in the Legal Interactive Task[1], we found that Indri search tool of Lemur 4.11, with its rich Boolean query language, is a very effective tool in utilizing the important keywords of a topic and retrieving a set of useful documents. So, we continued to use Indri for Boolean retrieval. To suit the nature of the Learning Task, we decided to use Terrier 3.0 to use its built-in relevant feedback support. In addition, we used our clustering algorithm (described in the following section) to maximize assessor feedback.

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

2.1 Clustering Algorithm

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 an arbitrarily chosen document as the representative (or seed) of the cluster for TA judgement. Such a cluster will be deemed relevant or nonrelevant according as its seed is judged as relevant or not.

2.2 Topic 401

The production request of topic 401 read as follows:

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

The coding instructions and the kick-off call provided crucial information about the notion of relevance of the topic. Based on this, we formed nine Indri queries:

1. #band(enrononline contact)
2. #band(enrononline guest access emails)
3. #band(enrononline #1(trade agreement))
4. #band(enrononline #syn(press news) release)
5. #band(enrononline #1(license agreement))
6. #band(enrononline #1(financial product))
7. #band(enrononline #5(trade derivative))
8. #band(enrononline financial #syn(instrument product derivative swap))
9. #band(enrononline #syn(marketing advertisement))

We submitted our first interim run - ISICLUT1 (prior to any assessment request) based on the documents retrieved by queries 1 and 2. We clustered all the documents returned by Indri queries 1 and 2 and manually inspected seeds from the clusters formed. All the seeds and their clusters deemed relevant were ranked above all other documents in the collection. We then sent the same seeds for assessor judgement. We continued our clustering exercise for the outputs of queries 3, 4, 5 and 6. By this time, we had finished our first 100 assessments. So, we submitted our second interim run - ISICLUT2. But now we had relevant document expanded from judged relevant seeds for queries 1 to 6. In this run, we arbitrarily ranked these documents and followed them with all other documents in the collection. Next we assessed all the seeds of 7, 8 and 9.

2.2.1 Relevance Feedback and Boolean Query Expansion

We had achieved the relevance judgements of the seeds generated by the nine Indri queries discussed shortly. Next we decided to use Rocchio relevance feedback technique implemented in Terrier 3.0. Using this technique, we applied query expansion based on the relevance information received hitherto. At this stage, we tried out expansion of Boolean Indri queries. We could have done this kind of expansion in two ways. One, we could have expanded each of the nine Indri queries constructed initially. Another variant is that, we could expand the basic query of 401 as a whole. We tried both the approaches.

Using only the relevant assessments of Indri query 1, we obtained eight Rocchio feedback terms, viz. contact, user, product, password, launch, gosalia, screen and amita. Here, the Terrier query contained the terms enrononline and contact. So, the expanded Indri query for query 1 is as follows:

#band(enrononline contact #syn(user product password launch gosalia screen amita))

Similarly, we obtained the following expanded version of query 3:

#band(enrononline #1(trade agreement) #syn(user ngpl product password screen click transact online offer))

We made only two querywise expansions, for queries 1 and 3, since they had high number of relevant documents which is necessary to generate good expansion terms.

We selected the new documents in top 100 of the ranked-list returned by Terrier after query expansion and sent them for TA assessment. For the new documents obtained by Indri expanded queries, we applied clustering and sent the seeds for assessment.

Next, we used all the relevant terms for 401 and obtained the following expanded Indri query:

#band(#syn(enrononline eol) #syn(tagg jarnold orig dollar bankruptcy may01 kiodex active nature)), where the basic Terrier query had the terms enrononline and eol.

2.2.2 Final Submission

At this stage, we had the relevant assessments of all the nine Indri queries and also those obtained through relevance feedback. We submitted the following three runs as the final submissions :

1. ISIROTTF : In this run, we placed the relevant documents in arbitrary order followed by all other documents in the collection.
2. ISITRFTF : We used Rocchio relevance feedback using all the relevant documents, the Terrier query containing the terms enrononline and eol
3. ISILRFTF : Here, the expansion terms returned by Rocchio feedback are used to expand the Indri query

   #band(#syn(enrononline eol)

2.2.3 Mop-up Submission

Before the mop-up submission, the relevance judgements of the documents submitted to the TA for assessment were released. We submitted the following runs, based on this relevance information :

1. ISIRoTAM : We placed the relevant documents in arbitrary order followed by all other documents in the collection.
2. ISITrFAM : We used Rocchio relevance feedback using all the relevant documents, the Terrier query containing the terms enrononline and eol
3. ISIFuSAM : this was the Z-fusion of the above two runs, with equal weight on each

2.3 Topic 403

We used similar techniques for Topic 403. The production request for Topic 403 was:

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

Based on the kick-off call and coding guidelines, we formed the following seven Indri queries:

1. #band(#syn(oil gas) enron spill environment)
2. #band(#syn(oil gas) enron spill #syn(cleanup #1(clean up)))
3. #band(#syn(oil gas) #syn(leak spill) #syn(environment atmosphere surrounding))
4. #band(#syn(oil gas) #syn(spill emission) #syn(violation negligence) #syn(rule law regulation) #uw25(prevent damage))
5. #band(#syn(oil gas) pollution #syn(damage hazard harmful) #syn(environment atmosphere surrounding) #syn(air water sea river ocean))
6. #band(enron noise pollution)
7. #band(enron animal habitat #syn(harm damage kill))

We made our first interim submission ISICLST1 using the documents retrieved by all the above queries. We clustered them, manually judged the seeds and placed the judged relevant clusters in the ranked list above the other documents in the list.

2.3.1 Relevance Feedback and Boolean Query Expansion

Like Topic 401, here also we used Rocchio relevance feedback and expanded the Boolean Indri queries. We expanded query 1 into the following query:

#band(#syn(oil gas) enron spill environment #syn(transrede tg plaintiff bbpl approximate mmcmd pipelin gtb court bolivian))

Here Terrier query contained the terms oil, gas, enron, spill and environment. Rocchio feedback yielded the terms - transrede, tg, plaintiff, bbpl, approximate, mmcmd, pipelin, gtb, court and bolivia. Similarly, query 2 was expanded into following:

#band(#syn(oil gas) enron spill #syn(cleanup #1(clean up)) #syn(elektro court transrede appeal teixeira silva serec freire dpc tozzini))

Finally, we did expansion for the entire topic (using the whole set of relevance assessments) and the expanded query was:

#band(#syn(oil gas) enron environment #syn(transrede court cuiab tg plaintiff pipelin approxim bolivian gtb elektro))

2.3.2 Final Submission

Our final submission for Topic 403 consisted in three runs:

1. ISIROTTF : We placed the relevant documents in arbitrary order followed by all other documents in the collection.
2. ISITRFTF : We used Rocchio relevance feedback using all the relevant documents, the Terrier query containing the terms - oil, gas, enron and environment
3. ISILRFTF : Here, the expansion terms returned by Rocchio feedback are used to expand the Indri query for the entire topic

2.3.3 Mop-up Submission

Before the mop-up submission, the relevance judgements of the documents submitted to the TA for assessment were released. We submitted the following runs, based on this relevance information :

1. ISIROTAM : We placed the relevant documents in arbitrary order followed by all other documents in the collection.
2. ISITRFAM : We used Rocchio relevance feedback using all the relevant documents, the Terrier query containing the terms - oil, gas, enron and environment
3. ISIFUSAM : this was the Z-fusion of the above two runs, with equal weight on each

3 Results

The results presented in this section were declared as "Preliminary evaluation results". The evaluation measure used was "hypothetical F1".

Run type	Run id	Hypothetical F1
INTERIM RUNS	ISICLUT1	8.8%
	ISICLUT2	8.8%
FINAL RUNS	ISIROTTF	12.2%
	ISITRFTF	50.5%
	ISILRFTF	17.6%
MOP UP RUNS	ISIRoTAM	8.8%
	ISITrFAM	57.8%
	ISIFuSAM	14.1%

Table 1: Topic 401 - Performance of runs

Run type	Run id	Hypothetical F1
INTERIM RUNS	ISICLST1	11.0%
	ISIRFCT2	7.2%
FINAL RUNS	ISIRoTTF	9.0%
	ISITrFTF	13.9%
	ISILrFTF	4.6%
MOP UP RUNS	ISIROTAM	32.5%
	ISITRFAM	24.9%
	ISIFUSAM	34.0%

Table 2: Topic 403 - Performance of runs

	Hypothetical F1			
Topic no	Best	Median	Worst	Our Best
401	58.8%	28.0%	8.8%	57.8%
402	58.8%	13.1%	2.3%	-
403	72.0%	14.2%	3.1%	34.0%

Table 3: Overall performance of all submitted runs

Table 1 and 2 show our performance for Topics 401 and 403. Table 3 shows the overall performance of all the submitted runs. As expected, the performance got better as more relevant documents were received. Relevance feedback proved to be a very effective tool in capturing new relevant documents from smaller number of judged relevant ones.

4 Conclusion and Future work

In our last year's participation, we learned that our clustering technique was quite effective in retrieving relevant documents. This yielded us very high precision but at the expense of recall and consequently, F-measure too. This led us to believe that restricting to relevant clusters would not serve the basic purpose of legal retrieval, which hinges on high recall. We had to use the initial relevance information to seek for more relevant documents from the entire corpus. We thought that relevance feedback and query expansion would suit our goal. In addition to the state-of-the-art text query expansion techniques, we made a humble attempt in Boolean query expansion, which produced promising results. In our mop-up submission, we also used data fusion in the hope of boosting couple of our runs. On the whole, as compared to our previous year's performance, we found that we have managed to strike a balance between precision and recall (reflected by F-measure).

In our clustering technique, we have basically implemented the single-linkage clustering algorithm. We are looking to implement the other type of hierarchical clustering - complete-linkage clustering. Also, in our technique we have chosen a representative document from a cluster (i.e., a *seed*) arbitrarily. Instead of doing so, we will like to devise a method of choosing the best representative *seed* of a cluster.

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

[4] Jacques Savoy. Data fusion for effective european monolingual information retrieval. pages 233–244. In Proceedings of CLEF'2004., 2004.