1 Introduction

The TREC Legal track was introduced in TREC 2006 with the claimed purpose of to evaluate the efficacy of automated support for review and production of electronic records in the context of litigation, regulation and legislation. The TREC Legal track 2008 runs three tasks: (1) an automatic ad hoc task, (2) an automatic relevance feedback task, and (3) an interactive task. We have only taken part in the automatic ad hoc task of the TREC Legal track 2008, and focused on the following issues:

1. **Indexing**. The CDIP test collection is characterized by a large number of unique terms due to OCR mistakes. We have defined a term selection strategy to reduce the number of terms, as described in Section 2.

2. **Querying**. The analysis of the past TREC results for the Legal track showed that the best retrieval strategy basically returned a ranked list of the boolean retrieved documents. As a consequence, we have defined a strategy aimed to boost the score of documents satisfying the final negotiated boolean query. Furthermore, we defined a method for automatic construction of a weighted query from the request text, as reported in Section 3.

3. **Estimation of the K value**. We have used a query performance prediction approach to try to estimate K values. The query weighting model that we have adopted is described in Section 4.

Submitted runs and their evaluation are reported in Section 5.

2 Indexing

The IIT Complex Document Information Processing (CDIP) test collection is based on a snapshot of the Tobacco Master Settlement Agreement (MSA) sub-collection of the LTDL.
**Seventeenth Text REtrieval Conference (TREC 2008) held in Gaithersburg, Maryland, November 18-21, 2008. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).**
The CDIP collection presents a large number of unique terms due to OCR mistakes. For example, GOV2, one of the biggest test collections (426GB), contains about 50ML unique terms for 25ML of documents, whereas CDIP (57GB of text) contains almost twice as much: 95ML unique terms for 7ML documents.

We have used the Terrier (TERabyte RetriEyE) Information Retrieval platform [1] using the following indexing configuration:

\[
\begin{align*}
    \text{TrecDocTags.doctag} &= \text{record} \\
    \text{TrecDocTags.idtag} &= \text{tid} \\
    \text{TrecDocTags.process} &= \text{ot} \\
    \text{TrecDocTags.casesensitive} &= \text{false} \\
    \text{Termpipelines} &= \text{Stopwords, PorterStemmer}
\end{align*}
\]

With this configuration all meta-data were ignored. In addition, words containing more than 20 characters were excluded.

With such settings the generated index would have had 95ML terms with 14GB of index structures (direct, inverted and lexicon files). Therefore, we have removed from index all terms having a very low frequency (i.e. \( \leq 25 \)) producing a lexicon with "only" 9ML terms.

We have noticed that the presence of OCR errors, by altering both global and local statistics, affects the retrieval quality.

### 3 Retrieval

We have used a variation of the DFRee model, a parameter free IR model offered by the Terrier platform. With this model we have not had to tune any parameters, and we have focused only on the proposed methodology, by evaluating the gain in performance with respect to the baseline. We have extracted the most informative terms from the top-returned documents as performed in query expansion. In this expansion process, terms in the top-returned documents are weighted using a particular DFR term weighting model. During our experiment we have found that Bo1 (Bose-Einstein statistics) term weighting models best fit with this collection. We have also made tuning on the number of top-returned documents and number of expanded term; we have obtained the best performance considering 5 documents and 10 terms.

#### 3.1 Boolean re-rank

We have boosted the scores of all documents \( d_i \in B \), where \( B \) is the set of documents satisfying the final negotiated boolean query. The score \( s_i \) of a document \( d_i \in B \) was augmented by a value \( s \):

\[
    s_i' = s_i + s
\]

where \( s = s_k \) is the score of the \( k \)-th document and

\[
    k = \alpha \ast b
\]
is the number of documents in \(B\) and \(\alpha\) a parameter. Notice that \(s_k\) varies on each topic because \(b\) depends on the final negotiated boolean query. If \(\alpha = 0\) then the score \(s_0\) of the top-rank document is added to all \(d \in B\), and all these documents are shifted up in all top-most positions. Documents also keep the order computed by the DFR model. On the other hand when \(\alpha \to \infty\) then \(s_i \sim s_i\).
4.2 Query weighting model

Since we had not time to extract expanded query weights we simply used the inverse document frequency:

\[ D_i = \sum_{t \in q_i} - \log \left( \frac{n_t}{N} \right) \]

Where \( t \) is a term of a query \( q_i \), \( n_t \) is the number of relevant documents where the term \( t \) appears, \( N \) is the number of all the documents in the collection. The correlation factor between \( D_i \) and \( K_i \), with \( K_i \) the optimal value for the \( F_1 @ K \), was about 30% with TREC data of 2007, whereas an higher correlation there was between \( D_i \) and the estimated number of relevant documents (37%).

4.3 Normalization’s variants

To turn the coefficient \( D_i \) into an estimated \( K \) value we used the Z-Score [4]. For the new queries we have formulated three different models:

1. \( K_O = \left( 1 - \frac{D - \mu_{idf}}{\sigma_{idf}} \right) \cdot \mu_k \)
   where \( \mu_{idf} \) and \( \sigma_{idf} \) are the mean and the standard deviation of all \( D_i \) respectively and \( \mu_k \) is the mean of the optimal \( K \).
2. \( K_R = \left( 1 - \frac{D - \mu_{idf}}{\sigma_{idf}} \right) \cdot \mu_R \)
   where \( \mu_R \) is the mean of the cardinalities of relevant documents.
3. \( K_B = \left( 1 - \frac{D - \mu_{idf}}{\sigma_{idf}} \right) \cdot \mu_B \)
   where \( \mu_B \) is the mean of the cardinalities of \( B \).

5 Submitted runs and results

For the participation to the ad hoc task we have submitted the following runs:

- The run labeled as \( CTF\text{rtSk} \) represents the compulsory baseline. It is computed just using the request text field as query. The \( K \) values are computed using the first model described in Section 4.3.
- The run labeled as \( CTF\text{rtSkBr0} \) is computed just using the (typically one-sentence) request text field as query. The boolean re-rank strategy is applied with \( \alpha = 0.0 \). The \( K \) values are computed using the first model described in Section 4.3.
- The run labeled as \( CTF\text{ggeSkBr0} \) computed using the query elaboration method described in Section 3. The boolean re-rank strategy is applied with \( \alpha = 0.0 \). The \( K \) values are computed using the first model described in Section 4.3.
- The run labeled as \( CTF\text{gge4kBr0} \) is computed using the query elaboration method described in Section 3. The boolean re-rank strategy is applied with \( \alpha = 0.0 \). The \( K \) values are equal to 40,000 for each topic.
- The run labeled as CTFgge10kBr0 is computed using the query elaboration method described in Section 3. The boolean re-rank strategy is applied with $\alpha = 0.0$. The $K$ values are equal to 100,000 for each topic.
- The run labeled as CTFggeBkBr0 is computed using the query elaboration method described in Section 3. The boolean re-rank strategy is applied with $\alpha = 0.0$. The $K$ values are computed using the third model described in Section 4.3.
- The run labeled as CTFggeBkBr1 computed using the query elaboration method described in Section 3. The boolean re-rank strategy is applied with $\alpha = 1.0$. The $K$ values are computed using the third model described in Section 4.3.
- The run labeled as CTFggeRkBr0 computed using the query elaboration method described in Section 3. The boolean re-rank strategy is applied with $\alpha = 0.0$. The $K$ values are computed using the second model described in Section 4.3.

Results of submitted runs are shown in Table 1.

<table>
<thead>
<tr>
<th>Run label</th>
<th>$F_1$</th>
<th>$H$</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Best run of the ad hoc task ’08</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$est_F_1 @ K$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$est_F_1 @ R$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$est_H F_1 @ K_h$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$est_F_1 @ R_h$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best run of the ad hoc task ’08</td>
<td>0.2204</td>
<td>0.2458</td>
<td>0.1064</td>
</tr>
<tr>
<td>Median run of the ad hoc task ’08</td>
<td>0.1429</td>
<td>0.1823</td>
<td>0.0702</td>
</tr>
<tr>
<td>CTFrtSk (baseline)</td>
<td>0.1282</td>
<td>0.1736</td>
<td>0.0481</td>
</tr>
<tr>
<td>CTFrtSkBr0</td>
<td>0.1346</td>
<td>0.1822</td>
<td>0.0723</td>
</tr>
<tr>
<td>CTFggeSkBr0</td>
<td>0.1544</td>
<td>0.2152</td>
<td>0.0811</td>
</tr>
<tr>
<td>CTFgge4kBr0</td>
<td>0.1849</td>
<td>0.2152</td>
<td>0.0818</td>
</tr>
<tr>
<td>CTFgge10kBr0</td>
<td><strong>0.2160</strong></td>
<td>0.2152</td>
<td>0.0788</td>
</tr>
<tr>
<td>CTFggeBkBr0</td>
<td>0.1800</td>
<td>0.2152</td>
<td>0.0799</td>
</tr>
<tr>
<td>CTFggeBkBr1</td>
<td>0.1856</td>
<td><strong>0.2196</strong></td>
<td>0.0824</td>
</tr>
<tr>
<td>CTFggeRkBr0</td>
<td>0.1785</td>
<td>0.2152</td>
<td><strong>0.0947</strong></td>
</tr>
</tbody>
</table>

With respect to the main metric (i.e., estimated $F_1 @ K$) the evaluation of our baseline (CTFrtSk) is 0.1282. Six of the eight submitted runs are over the median of the ad hoc task of Legal track 2008.

Applying the boolean re-rank to the baseline (CTFrtSkBr0) we improve our performance of +4.99%. The introduction of the query elaboration method (CTFggeSkBr0) allows to improve our performance of +20.43% with respect to the baseline.

If we consider only those runs in which the $K$ value is computed on basis of the query complexity, the best improvement with respect to the baseline is 44.77% (CTFggeBkBr1). This is not our best result. In fact the best run is CTFgge10kBr0. It improves the baseline performance of 68.48% considering $K=10000$ for each topic. CTFgge10kBr0 differs only -2.03% from the best run among all 64 runs submitted to the ad hoc task of Legal track 2008.
With respect to the estimated $H F_1@K_h$ metric the baseline is evaluated 0.0481. Seven of the eight submitted runs are over the median of the ad hoc task of Legal track 2008. Our best run is CTFggeRkBr0 with an evaluation of 0.0947 and an improvement of 96.88% over the baseline. CTFggeRkBr0 differs 12.35% from the best run among all 64 runs submitted for the ad hoc task of Legal track 2008.

We have already started a experimentation aimed to better analyze the relationship between the $D_t$ values, introduced in Section 4.2, and the retrieval performance.

References