The Robert Gordon University's HARD Track Experiments at TREC 2004

David J Harper, Gheorghe Muresan\textsuperscript{1}, Bicheng Liu, Ivan Koychev, Dietrich Wettschereck, and Nirmalie Wiratunga
School of Computing, The Robert Gordon University,
St Andrew Street, Aberdeen, UK.
(email: david.harper@smartweb.rgu.ac.uk)

Introduction and Motivation

This was the first time that RGU had participated in the HARD track, and indeed in TREC. We were interested in investigating the effect of exploiting the topic metadata to re-rank our initial baseline run, in a similar fashion to that of Rutgers in TREC 2003 [Belkin et al, 2003]. We used the Lemur toolkit (LTK) to obtain a baseline ranking, using title and description for each topic, and using OKAPI BM25 weighting (with default LTK settings). Then, we focussed on re-ranking this baseline for each topic, based on queries generated specifically to rank separately by genre, geography, and familiarity, using the LTK re-ranking capability (ranking of so-called “working set”). The baseline and metadata-derived rankings were then combined using an evidence combination approach.

Our experiments were motivated by several interests. First, we were interesting in re-ranking the topicality-derived baseline based on queries generated specifically for each meta-category/metadata pair. This enabled us to investigate the effect of each meta-pair source individually, and better understand the effectiveness of the approach used for that meta-pair. Moreover, the separate rankings provided a good basis for our subsequent approaches to evidence combination, in which we combine the various sources of evidence from both the baseline ranking and meta-pair re-rankings. (We will refer to the pair meta-category/metadata as a meta-pair.)

Second, we were interested in a variety of approaches for generating queries based on the various meta-pair specifications. We explored machine learning approaches, and specifically the use of relevance feedback, based on the training data. We also generated manual queries for some meta-pairs. And, we devised a novel topic-specific approach based on language modelling and the clarity measure, for ranking documents by familiarity.

Third, we wanted to explore a number of principled approaches to combining the evidence provided by the baseline and metadata-derived rankings. These included Dempster-Shafer evidence combination, and a fusion technique based on normalising scores across rankings using rank position.

Fourth, we were interested in the challenge of evaluating, and understanding, the potentially complex interactions between the various approaches we used. Specifically, we were interested in evaluating the individual effects of the various approaches used for metadata-based re-ranking, and the overall effect of evidence combination.

Ranking using Metadata

Three basic approaches were considered in re-ranking the topicality-based initial runs using the metadata provided for each topic\textsuperscript{2}, these being:

\begin{itemize}
\item Relevance feedback approach using relevance assessments from the training data;
\item Manual generation of queries for (some of) the meta-pair combinations; and
\item A novel approach for generating familiarity-specific queries based on building topic models for sets for pseudo-relevant documents from the baseline, and selecting terms based on the clarity measure.
\end{itemize}

\textsuperscript{1} Gheorghe Muresan, Rutgers University, was a visiting researcher at the Smart Web Technologies Centre, The Robert Gordon University, and contributed to the RGU TREC 2004 experiments.

\textsuperscript{2} We refer to “topic” when referring to the TREC HARD topics, and “topicality” when referring to the use of topic title, description and narrative in retrieval (by topicality).
Relevance Feedback Approach

The track participants were provided with training data comprising 21 topics, including both topicality and metadata specifications. For each topic, 100 documents were assessed for relevance. Document were assigned one of three labels (i) not “on topic”, (ii) “on topic” but not meeting metadata specification, or (iii) relevant. We will refer to these three kinds of document as “not relevant”, “soft relevant” and “hard relevant” respectively. In the case of the soft relevant documents, they were further labelled with the meta-categories they did not satisfy. As a result, we were able to produce sample sets of hard and soft relevant document for each meta-pair.

We conjectured that, given samples of hard and soft relevant documents for any meta-pair, we should be able to improve the baseline for topics with that specification, by using the samples in a relevance feedback process. The effectiveness of this process will depend on at least two factors. We need to obtain a large enough sample of hard relevant documents for each meta-pair. And, given that the sample is derived from a number of topics, we need to ensure that the sample is not biased towards particular topics (i.e. topicality-biased). Alternatively, we need ways of factoring topicality from metadata-ness in each sample. Given the small number of topics and training documents, we did not expect to satisfy these constraints.

Manual Query Generation

This approach was effectively a fallback position, if the relevance feedback approach did not work. We generated a set of working conjectures for manual generation of queries, based in part on the characteristics of the training data, and in part on the characteristics of the corpus. These conjectures relate to the individual meta-category/metadata pairs, and we describe both the conjecture and our approach to query generation:

Genre/news Given the corpus is a news corpus, and thus dominated by news-type stories, it is highly likely that this specification would be met almost by default. We felt that in effect genre/news might be treated by the users as if it were genre/any! Therefore, we decided not to generate queries of this type.

Genre/oped Given this requires a very specific kind of document (opinion/editorial), it is likely that the user would satisfy themselves that this criteria was met. We generated a manual query based on inspection of the small sample of genre/oped documents in the training data. We focussed on words expressing the “first” person, opinions and views, and topical words typical of reviews, e.g. book, film, etc.

Genre/other We thought that “other” would be applied to very specialised or technical topics, and conjectured that the topicality parts of the topic might prove sufficient in retrieving appropriate documents. Therefore, we did not generate a manual query.

Geography/(US and non-US) For the geography specification, we conjectured that US (resp. non-US) geography could be approximated using a query comprising US (resp. non-US) places names. We generated two queries using the names of US states, state capitals, and state mnemonics for the “US” query, and country names and “nationality” for the non-US query, (e.g. “Iran, Iranian/ France, French, etc.).

Familiarity/(little and much) We did not generate manual queries as we developed an automatic procedure for generating familiarity-specific queries (see below).

Familiarity Ranking using Clarity-based Approach

We formulated the following hypothesis concerning familiarity:

Users unfamiliar with a topic will prefer documents in which representative terms occur, and users familiar with a topic will prefer documents in which highly discriminating terms occur.
If we can identify these sets of representative and discriminating terms for a topic, then the term sets could be used as a query to re-rank the baseline according to familiarity.

We operationalised this hypothesis on the basis of the clarity measure. The top-ranking $K$ documents for each topic from the baseline were assumed relevant (c.f. pseudo relevance feedback), and the terms in these pseudo-relevant documents were ranked by the clarity measure [Cronen-Townsend et al., 2002]. We selected the top-ranked $L$ terms for further processing. The clarity measure is given by $\log (p_{R} / p_{C})$, where $p_{R} =$Prob(termi $|$ pseudo-relevant) and $p_{C} =$Prob(termi $|$ Collection). We refer to the first term $p_{R}$ as clarity-representation, and the second term $\log (p_{R} / p_{C})$ as clarity-discrimination, and we further rank the $L$ terms by these measures separately. We then selected the top-ranked $M$ terms from the clarity-representation terms as a query for re-ranking the baseline for those topics with familiarity specification ‘little’, and similarly from the discrimination-clarity terms for ranking according to the familiarity specification ‘much’.

Evidently, this procedure is heavily dependent on the whether an adequate sample of “on topic” (they may indeed be soft or hard relevant) documents is obtained from the baseline. We know the pseudo-relevance feedback process can actively harm retrieval if this sample is poor in relevant documents. It is likely that for topics that perform poorly in the baseline run, our approach may not improve the baseline with respect to familiarity, and indeed may harm it.

The familiarity-specific queries are purposefully topic-specific, and hence good queries may improve the baseline in two ways. First, they may improve the soft effectiveness, by promoting soft relevant documents in the re-ranking. Second, if our conjecture holds, they may improve hard effectiveness, by promoting hard relevant documents in the re-ranking.

We also conjectured that the familiarity/much specification is likely to be considered more important by the users than the familiarity/little specification. By their nature, news articles (in the general sense of all kinds of news output), are written on the assumption that the reader will indeed have little familiarity with any given topic. Therefore, we believe that most documents will meet the familiarity/little specification almost by default.

Evidence Combination

Before describing our approaches to evidence combination, we wish to discuss ranking-based versus filter-based use of metadata. In our ranking approach, we use each source of evidence for the meta-pairs to generate a separate re-ranking of the baseline. In a filter-based approach, one might use a given source of evidence to filter (i.e. remove) documents from the baseline ranking, e.g. filter baseline based on geography/US (say). We prefer the ranking approach for two reasons. First, we want to preserve any data obtained from a source of evidence for as long as possible in our process, and essentially let the evidence combination deal with poor scoring documents. The risk with filtering is that a document may score highly based on most sources, and be removed based on a poor source of evidence. This is particularly an issue when some of our approaches to metadata are very speculative, or indeed very simplistic! Second, we wanted to explore principled approaches to evidence combination (or fusion), and in part this means that all sources of evidence should be considered in toto.

In previously reported work on the HARD track, metadata evidence was used to adjust the baseline ranking in a relatively ad hoc way [Belkin et al, 2003]. Thus, baseline scores were adjusted based on heuristics developed for each sources of evidence, e.g. readability scores used to rank by Flesch Reading Ease Score. In this work, we wished to explore principled approaches in which each sources of evidence was treated alike, although we hasten to add not necessarily with equal weight.

We explored two approaches to evidence combination, namely Dempster-Shafer evidence combination, and an approach based on normalising scores based on rank position, and weighted combination of the resultant normalised scores.

---

3 Technically, this is the contribution to the clarity measure based on the Kullback-Leibler divergence between the topic model and the collection model.

4 http://csep.psyc.memphis.edu/cohmetrix/readabilityresearch.htm
Dempster-Shafer This approach was first proposed for use in an IR context by [Jose and Harper, 1997] for combining sources of evidence for image retrieval, and has subsequently been used for in other image retrieval research [Aslandogan and Yu, 2000], and in [Urban et al, 2003].

We will not describe Dempster-Shafer evidence combination in detail, but rather give the reader a flavour for how it can be applied in document retrieval. Essentially, each source of evidence (e.g. a set of document scores in a given ranking) can be viewed as providing support for so-called singleton sets comprising a set containing each individual document. In Dempster-Shafer, for each source of evidence, we have a confidence level between zero and one for each source. A confidence level of one means we have complete confidence in that source, and zero that we have no confidence. Suppose for a given source, we have confidence \( C \). Further, each source of evidence has a base probability assignment (BPA) or mass, which in our application is a set of normalised scores summing to \((1-C)\) for the documents in a given ranking. Strictly, \( C \) is that part of the BPA that is unassigned to any proposition set.

Let us assume we have generated the BPAs for two sources A and B, as follows, and we have confidence \( C_A \) and \( C_B \) in these sources as shown:

\[
\begin{align*}
m_A &= (a_1, a_2, \ldots, a_n), \text{confidence } C_A \quad \text{[note } \sum a_i = 1 - C_A \text{]} \\
m_B &= (b_1, b_2, \ldots, b_n), \text{confidence } C_B \quad \text{[note } \sum b_i = 1 - C_B \text{]}
\end{align*}
\]

where \( n \) is the number of documents in the source/ranking, \( a_i \) is the score for document \( i \) according to source A, and similarly for \( b_i \) and source B.

Let \( m_A(i) \) denote \( a_i \) similarly \( m_B(i) \). The (simplified) rule for combining singleton sources of evidence to obtain the new BPA (or mass) is:

\[
m_{\text{Comb}}(i) = m_A(i) m_B(i) + (1-C_A) m_B(i) + (1-C_B) m_A(i), \text{ and}
\]

confidence \( C_{\text{Comb}} = 1 - (1-C_A) (1-C_B) \).

This combination rule has some very nice properties [Jose and Harper, 1997; Jose, 1998], when you explore the various limiting values. For example, when \( C_A=C_B=1 \), and we have complete confidence in both sources of evidence, then the rule shows we should multiply the scores (individual masses. On the other hand if \( C_A \) and \( C_B \) approach zero, then the rule shows we should add the scores. If we have no confidence in a particular source and complete confidence in the other (\( C_A=1, C_B=0 \), say), then the result is identical to considering source A by itself. Values between 0 and 1 provide “mixtures” of these kinds of behaviour.

There are potentially problems in applying Dempster-Shafer. First, scores derived from a variety of processes must be transformed into BPAs, and these BPAs should approximate to a probability distribution over the set of documents. Second, with large numbers of documents (\( n \) large), the individual masses become very small, and the multiplicative term in the combination is dominated by the additive terms.

**Weighted Score-Rank Method** For a given topic, we assume that we have ranked scores for each relevant sources of evidence. Thus, for a topic with metadata specification genre/any, geography/US, and familiarity/much, we would have the baseline ranking, and a re-ranking of this baseline corresponding to geography/US, and familiarity/much. Depending on the way these rankings were obtained, the range and distributions of scores can be very different, and normalising scores across the rankings becomes an issue. In this approach, we substitute scores based on rank position for the actual scores, which is one way of normalising these scores\(^6\).

\(^5\) In general, Dempster-Shafer enable one to combine evidence for sets of propositions, e.g. for a set of documents. But, for our purposes, it is sufficient to deal with individual documents.

\(^6\) Since submitting the TREC runs, Muresan has developed an approach to normalisation based on the use of z-scores, scores based on assuming a normal distribution over scores, and using standard deviation from the mean as the score.
Let us assume that the baseline ranking contains $T_{max}$ documents, then for a document at rank $j$ in a particular ranking, we compute a rank score of $(T_{max}+1-\text{rank}_j)$. This assigns the maximum score ($T_{max}$) to the top-ranked document, and for a document at rank $T_{max}$, a score of 1. If a document in the baseline is not retrieved for a given metadata source, then it is assigned a score of zero.

Each source of evidence is then allocated a weight. The score for a given document is computed as the weighted average of the rank-based scores for that document from the relevant sources of evidence. For the example above, the baseline, geography/US and familiarity/much rankings would contribute to a score for each document appearing in the baseline for the given topic. Clearly, different topics would combine different sources of evidence.

The weight for a source of evidence (i.e. ranking) is based on our confidence in the source. These weights could, in principle, be learnt from the training data, but the training data was quite sparse for some meta-pairs. Instead, we chose (guessed) the weights based on intuition. We considered the baseline (topicality-based) ranking to be of most importance to the putative user, and allocated this source a weight of one in all our runs. We weighted the other sources, based on our confidence in the sources, derived through inspection of the re-ranking of the baseline, and on our conjectures (see section Manual Query Generation) about the relative importance of the meta-pairs to the users.

**Experimental Setup**

In all reported experiments, we used the Lemur toolkit to generate the baseline and to perform the re-rankings of this baseline.

The initial baseline was obtained using the Lemur toolkit (LTK), and Okapi BM25 weights. We used the title and description for each topic, and retrieved 1000 document per topic.

To simplify the re-ranking experiments, we generated a mini-corpus comprising all documents in the baseline, plus the documents in the training data. This mini-corpus enabled us to efficiently run re-ranking experiments using the LTK ranking over “working set”, i.e. over the mini-corpus. We note that, in general, the mini-corpus would have very different statistical properties than the complete corpus, but for most experiments reported here, these differences can be ignored. We note those cases where the statistical properties may be relevant.

For the relevance feedback experiments, we used the KL method (with feedbackDocCount = 5, feedbackTermCount = 20 and the default parameter settings for mixture smoothing).

For the manually generated queries, we used the KL method to re-rank the entire mini-corpus using the generated queries. Thus, we obtained complete rankings of the mini-corpus for genre/oped, geography/US and geography/non-US, independent of topic (at this point). We then generated rankings on a per topic basis, using the scores extracted from the mini-corpus ranking, for each meta-pair. Note, that we only generated a ranking for a topic if the meta-pair was specified for that topic.

In generating topic-specific queries for familiarity based ranking, we generated queries for each topic, depending on the familiarity metadata specification. We constructed a language model over the top-ranked $K$ ($K=10$) for each topic, and a language model over the entire HARD track corpus. We then generated either a familiarity/little or familiarity/much query for each topic, depending on the metadata specification. We chose the top $L$ ($L=50$) terms from clarity ranking, generated the appropriate clarity-representation (resp. clarity-discrimination) rankings, and selected the top $M$ ($M=20$) terms for the little (resp. much) query.

We used the LTK KL method, and re-ranked the mini-corpus using the appropriate query on a per topic basis. Thus, we obtained a re-ranking of the baseline for the Familiarity/little topics, and similarly for the Familiarity/much topics.

The evidence combination experiments were leveraged of the original baseline in all cases. For each topic, the appropriate re-ranked results were combined to obtain a new overall ranking of the baseline. Thus, for a topic with metadata specification genre/any, geography/US, and familiarity/much, we
would have the baseline ranking, combined with re-rankings of this baseline based on geography/US, and familiarity/much.

**Experimental Results**

First, we present some results about the properties of the collection, and specifically the HARD training topics and evaluation topics. The statistics on the training topics resulted in us deciding not to use the relevance feedback approach in the official runs we submitted. Then, we present the overall document-level results for the various official runs we submitted, and sketch some initial conclusions. We then present a topic-by-topic analysis that suggests that some of the meta-pair sources may be improving hard effectiveness. Finally, we present some preliminary data on the individual performance of each meta-pair source.

Before presenting the results, we note that all runs are based on re-ranking the originally submitted baseline, and thus the effects of the various approaches are limited to re-ranking of this baseline.

**Ranking using Metadata**

**Relevance Feedback Approach** In Table 1, we present summary data for the training topics. It would seem that there are only three meta-pairs, for which the training data might provide a reasonable sample of hard relevant documents, namely: genre/news, geography/US, and familiarity/little. There are a comparatively large number of positive training instances, and a range of topics represented. We believe that the training data for geography/US is unlikely to capture the concept of US-ness. It may be that good models could be derived for ‘news’ and ‘little’ through (positive) relevance feedback. However, we did not pursue the idea of using relevance feedback further, at least in the evaluation runs.

<table>
<thead>
<tr>
<th>MetaCat</th>
<th>MetaData</th>
<th>Positives</th>
<th>% Pos</th>
<th>Negatives</th>
<th>% Negs</th>
<th># topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genre</td>
<td>news</td>
<td>241</td>
<td>94.9</td>
<td>13</td>
<td>5.1</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>oped</td>
<td>2</td>
<td>25</td>
<td>6</td>
<td>75</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>18</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>any</td>
<td>76</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Geography</td>
<td>US</td>
<td>103</td>
<td>74.1</td>
<td>36</td>
<td>25.9</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>non-US</td>
<td>85</td>
<td>89.5</td>
<td>10</td>
<td>10.5</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>any</td>
<td>149</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Familiarity</td>
<td>little</td>
<td>288</td>
<td>99.7</td>
<td>1</td>
<td>0.3</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>much</td>
<td>49</td>
<td>98</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

**Evidence Combination**

**Dempster-Shafer** In applying the Dempster-Shafer method, in the way described in this paper, we need to normalise the scores for the different sources of evidence. Essentially, we need to transform the raw document scores for each source into a basic probability assignment (BPA). The obvious way to do this is to simply divide each score by the sum of all scores present in the ranking. There are however two problems with this solution. First, the resultant probabilities are extremely small for re-rankings contained 1000 documents. As a result, if you combine them using the combination rule, essentially the scores will be added together. Secondly, and more problematically, in the re-rankings derived from the manually generated queries, and to a lesser extent, the familiarity queries, the range and distribution of documents scores is extremely skewed. Consequently, most of the mass in the BPA will be attributed to just a few documents. We believe that solutions to these problems can be found, but given these difficulties, we decided to focus our efforts on the weighted score-rank method.
Weighted Score-Rank Method  We used this method to combine the evidence from the baseline run, and the metadata derived rankings. The details of each run we submitted are summarised in Table 2.  As indicated earlier, the original baseline is assigned a weight of 1.0, with smaller weights for the arguably less reliable/less important meta-pair rankings.

In run 10, we included two sources based on topicality, the original baseline, and a re-ranking of the baseline using the Lemur KL method with pseudo relevance feedback. We weighted these equally. Run 10* is a notional run, which is equivalent in effect to Run 10, and introduced for discussion purposes. It weights the topicality at 1.0 in combination, and shows the comparatively lower contributions of the metadata sources.

Clearly, it is difficult with the runs as submitted, to separate out the effects of the various metadata sources on performance. Later, we present the results on other runs, in which we explore the effect of each metadata source individually.

Table 2:

<table>
<thead>
<tr>
<th>Source Metadata Category</th>
<th>Metadata</th>
<th>Topicality</th>
<th>Run 1</th>
<th>Run 5</th>
<th>Run 6</th>
<th>Run 10</th>
<th>Run 10*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genre</td>
<td>Opinion - U.S</td>
<td></td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>Geography</td>
<td>Editorial U.S.</td>
<td></td>
<td>0.2</td>
<td>0.5</td>
<td>0.6</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>Familiarity</td>
<td>Little Non-U.S.</td>
<td></td>
<td>0.2</td>
<td>0.5</td>
<td>0.4</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>Much</td>
<td></td>
<td>0.5</td>
<td>0.5</td>
<td>0.6</td>
<td>0.6</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Table 3 shows the document-level results for the runs submitted that are based on the weighted score-rank method. The italicised results are those pertaining to the original baseline. For each run, we report average precision, relevant retrieved at rank 10, and R-precision, and we will focus mainly on R-precision in this discussion. Note, that the hard results are based on hard relevant assessments (on topic document meeting metadata specification in topic), and the soft results are based on soft relevant assessments (on topic). In general, hard and soft results should not be compared against each other. Given that the data values have large standard deviations, and are not normally distributed, we use two non-parametric tests to test statistical significance. They are the Wilcoxon Signed Ranks test, and the Sign test, both applied at significance level 0.05. Each significance test was performed against the baseline entry in the same column, and is only computed for the hard results. Figures marked with one/two asterisks show figures that are significantly different from the baseline, based on the Wilcoxon and Sign test respectively.

In respect of Hard R-Precision and Rels@10, all metadata runs had higher average R-precision than the baseline, and especially so for Run 10. However, most of these improvements are not statistically significant. But, the R-Precision result is significant for Run 1 (Sign), and run 10 (Wilcoxon and Sign). For run 1, there are respectively 21/9/16 topics for which R-Precision is better/worse/tied. For run 10, the corresponding figures are 23/9/14 topics.

Let us look at the results in more detail. Run 10 is clearly best when considering the hard results. Potentially, this improvement may be due to at least two factors. It may be due to improvements in soft effectiveness due to the inclusion of the second topicality-based source. And, it may be due to improvements in hard effectiveness due to the influence of the metadata ranking. It is instructive to look at the various metadata runs to try and understand what is going on.

Any difference between run6 and run10 is due to the inclusion of an additional “baseline” run, which is a pseudo relevance feedback run based on the original OKAPI baseline. For the soft results, we observe that for R-precision the difference is 4% points, and for Rels@10 8% points. For the hard
results, the improvement in percentage points is respectively 13% and 11%. It would appear that
exploiting the metadata may be providing an additional performance boost over that achieved through
simply adding the additional “baseline” run.

Comparing run1 and run10 is also instructive. The soft results are effectively the same for these two
runs. But, the run10 (or equivalently run10*), the hard Rel@10 and R-precision values are 6 and 8
percentage points higher than for run1. If we examine the weightings given to the metadata sources,
they are lower for run10* (same effective weighting as submitted run10) than for run1, and we attribute
this is achieving better balance between the topicality-based and metadata-based evidence.

Table 3: Document-level evaluation of all submitted runs. Standard deviation in (brackets);
percentage change (%) compared with baseline below that. */** indicate statistical significance at level
0.05 using Wilcoxon/Sign test resp.

<table>
<thead>
<tr>
<th>RUN</th>
<th>HARD</th>
<th>SOFT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avg Prec.</td>
<td>Rel@10</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.259</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>0.251</td>
<td>3.11</td>
</tr>
<tr>
<td></td>
<td>(0.240)</td>
<td>(0.246)</td>
</tr>
<tr>
<td>5</td>
<td>0.2432</td>
<td>0.261</td>
</tr>
<tr>
<td></td>
<td>(0.242)</td>
<td>(0.249)</td>
</tr>
<tr>
<td>6</td>
<td>0.224</td>
<td>0.251</td>
</tr>
<tr>
<td></td>
<td>(0.223)</td>
<td>(0.243)</td>
</tr>
<tr>
<td>10/10*</td>
<td>0.258</td>
<td>0.261</td>
</tr>
<tr>
<td></td>
<td>(0.240)</td>
<td>(0.250)</td>
</tr>
</tbody>
</table>

These results are very encouraging for three reasons. First, the weightings for the different evidence
sources have not been optimised, and were simply best guesses. Second, our treatment the metadata
categories genre and geography was extremely simplistic, and in fact, we ignored the ‘news’ and
‘other’ genres completely. (Note: this may have been advantageous as effectively genre/news did not
appear to be a strong factor in the user assessments.) Third, the initial baseline was rather low, and our
familiarity approach assumes that the top-ranked documents in the baseline provide a good sample of
relevant documents, c.f. pseudo relevance feedback. Further analysis is required in order to attribute the
observed performance improvements in run 1 and run 10 to particular sources.

Exploratory analysis of individual sources

We ran a series of experiments in which we combined the baseline source with a single other evidence
source, and measured the effect of the second source. We used the weighted score-ranks approach with
the baseline weighted at 1.0, and the second source weighted variously as shown in Table 4. In the
case of a given meta-pair source, we only explored the effect for those topics with that specified meta-
pair. Table 4 summarises the experiments, and reports the performance for R-Precision only. Again,
we test statistical significant using the Wilcoxon and Sign tests, comparing the combination against the
baseline performance within a particular row of the table.

The familiarity/much source improves the baseline significantly, and to a lesser extent so does the
familiarity/little source. It would seem that our new clarity-based approach to familiarity ranking is
highly effective, and particularly as it is based on pseudo-relevant documents from the baseline.
Neither of the geographic sources (US or non-US) improves the baseline, and indeed the non-US
source harms effectiveness. The reason for the poor non-US source performance is likely due to the
original topics. In all cases, the topics with non-US specification, included geographic places names in
the title/description, which means the baseline already including specific geographic boosting. The
relatively high baseline performance (R-Precision, 0.42) attests to this. Our more generic non-US
query was then highly likely to damage this already good (hard) baseline performance. Combining the
original OKAPI BM25-based ranking with a re-ranking based on LEMUR/KL with pseudo relevance
feedback, significantly increased R-precision. It may be that the very different natures of the retrieval mechanism each contributed something different, and together resulting in an improvement. This phenomenon has been observed generally in fusion/evidence combination.

It will be interesting to re-run the combination runs, with weightings adjusted to reflect more accurately the effectives of the individual sources of evidence.

Table 4: Document-level evaluation of baseline combined with a single source using the weighted score-ranks approach. */** indicate statistical significance at level 0.05 using Wilcoxon/Sign test resp. Bold results best for row. All results evaluated using hard relevance.

<table>
<thead>
<tr>
<th>Second source</th>
<th># topics</th>
<th>Base R-Pr</th>
<th>Second source weights and R-Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R-Pr</td>
<td>wgt R-Pr</td>
</tr>
<tr>
<td>Genre/oped</td>
<td>7</td>
<td>0.165</td>
<td>0.5</td>
</tr>
<tr>
<td>Geog/US</td>
<td>14</td>
<td>0.132</td>
<td>0.5</td>
</tr>
<tr>
<td>Geog/Non-US</td>
<td>8</td>
<td><strong>0.420</strong></td>
<td>0.5</td>
</tr>
<tr>
<td>Fam/little</td>
<td>25</td>
<td>0.176</td>
<td>0.5</td>
</tr>
<tr>
<td>Fam/much</td>
<td>20</td>
<td>0.342</td>
<td>0.5</td>
</tr>
<tr>
<td>Rerank Base</td>
<td>45</td>
<td>0.250</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Conclusions

In relation to our treatment of the metadata sources of evidence, we conclude that:

- Our new clarity-based approach to generating queries for familiarity (much and little) was highly effective, and significantly so for the familiarity/much queries. We note that these queries were topic-specific and not generic, and it is possible that the improvements are due to increases in performance of both soft and hard ranking. Importantly, our approach did not use any training data in generating the queries, although it did use pseudo-relevant documents from the baseline.

- The generic geographic queries were not effective in improving performance, although in the case of the non-US topics, this may be partly attributable to the innate geographic-bias of the topicality parts of the topic.

- The genre/opinion-editorial manual query surprisingly improved performance but not significantly so.

- Although technically not a metadata source, the inclusion of a second topicality sources, based on re-ranking the baseline using a different retrieval mechanism, proved highly effective, and significantly so.

In relation to our evidence combination approach, and our general approach to ranking based on each source of evidence, and combining the rankings, we conclude that:

- The weighted score-rank approach proved effective in combining very different sources of evidence, and significantly so in the case of our submitted run 10.

- Our evidence combination approach enabled us to systematically explore the individual effects of the various metadata sources, which results in more insights concerning the effectiveness of each source.

---

7 This result was close to being statistically significant.
8 Also t-test significant at p<0.032
In relation to experimental methodology, we offer the following observations:

- Our approach to ranking sources separately, and subsequent combination, proved useful in understanding the effects of the various metadata sources.
- By performing a more detailed analysis of the results beyond the overall run averages, we were better able to understand why some of our metadata approaches were effective or otherwise;
- The paired significance tests, and particularly the use of non-parametric Wilcoxon and Sign Ranks tests, were useful in discovering differences in data, which generally was highly correlated, and had large standard deviations.

Acknowledgements

This work was undertaken within the Smart Web Technologies Centre that was established through a Scottish Higher Education Funding Council (SHEFC) Research Development Grant (No. HR01007). Gheorghe Muresan and David Harper were supported through the Research Development Initiative of The Robert Gordon University. We would like to thank Dianne Kelly for her feedback on this work.

References