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

David J Harper, Gheorghe Muresan[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)

1. Introduction and Motivation

The High Accuracy Retrieval from Documents (HARD) track explores methods of improving the accuracy of document retrieval systems. As part of this track, the participants have investigated how information about a searcher's context can be used to improve retrieval performance [Allan, 2003; Allan, 2004]. Searchers, referred to as assessors in this track, produce TREC-style search topics. Additionally, assessors are asked to specify values from pre-defined categories of metadata, that relate to the context of the search, as opposed to the topic of the search. The categories and values of the metadata used for the HARD track in 2004 are:

- Desired genre of documents, with values *news*, *opinion-editorial*, *other*, and *any*.
- Desired geographic treatment, with values *USA*, *non-USA*, and *any*. (Documents satisfying the *USA* metadata value may also refer to non-USA geographic areas, and vice versa.)
- Assessor's familiarity with the topic, with values *much* and *little* (or equivalently *high* and *low*).
- One or more documents related to the topic, but not from the collection to be searched.
- Desired granularity of response, with values *passage* and *document*.

The work reported in this paper focuses on exploiting the genre, geographic and familiarity metadata. We refer to the combination of metadata category and value, e.g. genre/news, as a meta-pair.

The experimental protocol for the track is described in the overview paper [Allan 2004]. Track participants were provided with a set of twenty training topics, including the associated metadata relating to context, plus 100 training documents per topic and relevance assessments. The training documents were retrieved from the HARD 2004 test corpus, derived from a number of different news databases. Each document is assessed as being not relevant, soft relevant or relevant (which we refer to as hard relevant). Not relevant means the document is not on topic. Soft relevant means a document is on topic but does not meet all the metadata specified; hard relevant (or simply relevant) means a document is both on topic and meets the metadata specified. For soft relevant documents, the metadata categories they did not satisfy were also indicated.

After a training period, fifty unseen test topics were distributed, without the metadata. Participants then searched the test corpus using their local systems, and submitted baseline search results, 1000 retrieved documents per topic. After submission of the baselines, the metadata for the test topics was distributed. Participants could use this metadata in any way in order to produce a new set of final search results, which were then submitted for evaluation. Both baseline and final search results were evaluated by the assessors for each topic, using three-valued relevance assessment. The effectiveness of the use of metadata was determined by comparing the baseline search results with final search results, especially with respect to difference in hard relevance, using a variety of TREC measures. The HARD track gave prominence to R-Precision, which emphasizes high accuracy retrieval.

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,

[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.

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-pair, i.e. pair of metadata category/metadata value. 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.

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 Kullback-Leibler divergence, 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.

2. Ranking using Metadata

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

- Relevance feedback approach using relevance assessments from the training data;
- Manual generation of queries for (some of) the meta-pair combinations; and
- 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 computing the Kullback-Leibler divergence between a topic model and the overall collection model.

2.1 Relevance Feedback Approach

Using the training data, we were able to produce sample sets of hard and soft relevant documents for each meta-pair. For example, from those topics with meta-pair *genre/news*, we were able to generate a sample of hard relevant documents, and a sample of soft relevant documents. We conjectured that, by using the meta-pair samples in a relevance feedback process, we should be able to improve the baseline for topics with that specification. 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.

[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).

2.2 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).

2.3 Familiarity Ranking using Topic Specificity or Generality

We conjecture that people with low familiarity with a topic will prefer documents which are in general representative of the topic as a whole, whereas people with high familiarity with a topic will prefer documents which are specific to particular aspects of the topic. This is related to the intuition, and common experience, that people with little knowledge are better off reading something general about a topic, rather than something very specific. We attempted to approximate these characteristics by identifying documents which contain terms which are representative of a topic for users with low familiarity, and documents in which highly discriminative terms occur for users with high familiarity. Identifying such terms, with respect to any specific topic, gives us a set of terms which can be used as a query, for re-ranking the baseline search results. We thus formulated the following hypothesis concerning familiarity:

> *Users unfamiliar with a topic will prefer documents in which* ***highly 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 used the divergence between the collection language model and a topic language model as the basis for identifying representative and discriminating terms as follows. We define the collection model to be the probability distribution $P(t|C)$, where for every term t in the vocabulary, the model gives the probability that we would observe term t if we randomly sampled from the collection C. A topic model, $P(t|T)$, is constructed using the top-ranking K documents from the baseline, which are assumed to be relevant (pseudo relevant). The Kullback-Leibler divergence has been used extensively in applications of language modelling in information retrieval [Croft and Lafferty 2003; Lafferty and Zhai 2001]. In this approach, we compute the Kullback-Leibler divergence between a topic model and the collection model, and we restrict the computation to the terms occurring in the topic model, *VT*, as follows:

$$KL\,(T \parallel C) = \sum_{t_i \in VT} p(t_i \mid T) \log\left(p(t_i \mid T) \middle/ p(t_i \mid C) \right)$$

We ranked the terms in *VT* by their individual contribution to *KL (T || C)*, and selected the top-ranked *L* terms for further processing. We refer to the expression before the *log* as the **KL-representation**, as this gives a measure of the term density, and thus generality. We refer to the *log* expression as the **KL-discrimination**, as this gives a measure of term discrimination, and thus specificity.

Finally, the familiarity hypothesis is operationalised in the following way. KL-representative and KL-discriminative terms for each topic were identified by constructing a language model (using Lemur) for the first ten documents in the baseline results, and a language model for the entire HARD corpus. Then, the top 50 terms contributing to *KL (T || C)* were selected, and the KL-representation and KL-discrimination expressions were computed for those terms. For each topic with familiarity level *low* (or *little*), the 20 top-ranked KL-representation terms were selected to construct a query which was then used to re-rank the baseline results for that topic. For each topic with familiarity level *high* (or *much*), the 20 top-ranked KL-discrimination terms were chosen to construct a query used to re-rank the baseline results for those topics.

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.

3. 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*[3]. In this work, we wished to explore principled approaches in which each source of evidence was treated alike, although we 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.

3.1 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[4] 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 sources, 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 C for the documents in a given ranking. Strictly, $(1-C)$ 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:

$m_A = (a_1, a_2, \ldots., a_n)$, confidence C_A $\quad , \sum a_i = C_A$

$m_B = (b_1, b_2, \ldots., b_n)$, confidence C_B $\quad , \sum b_i = C_B$

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:

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

confidence $C_{Comb} = C_A + C_B – C_A\ C_B$

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

[3] http://csep.psyc.memphis.edu/cohmetrix/readabilityresearch.htm

[4] 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.

individual masses become very small, and the multiplicative term in the combination is dominated by the additive terms.

3.2 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[5].

Let us assume that the baseline ranking contains *Tmax* documents, then for a document at rank *j* in a particular ranking, we compute a rank score of *(Tmax+1-j)*. This assigns the maximum score *(Tmax)* to the top-ranked document, and for a document at rank *Tmax*, 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 2.2, Manual Query Generation) about the relative importance of the meta-pairs to the users.

4. 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.

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

For the manually generated queries, we used the LTK KL method to re-rank the entire mini-corpus using the generated queries. Thus, we obtained complete rankings of the mini-corpus for the meta-pairs *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

[5] 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.

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 the KL ranking (see section 2.3), generated the appropriate KL-representation (resp. KL-discrimination) expressions, and selected the top-ranked *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 *familarity/little* topics, and similarly for the *familiarity/much* topics.

The evidence combination experiments were leveraged off 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*.

5. 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.

5.1 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 1: Summary data for training topics: s of documents judged hard (positive) and soft (negative) relevant. Number of topics contributing the data is also given.

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

5.2 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 sources 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: Details of evaluation runs showing weighting of difference evidence sources used in applying the weighted score-rank method.

Source	Metadata					Topicality	
Metadata Category	Genre	Geography		Familiarity		Baseline: Lemur/Okapi	Reranked baseline: Lemur/KL with pseudo RF
Metadata Value	Opinion-Editorial	U.S.	Non-U.S.	Little	Much		
Run							
1	0.4	0.2	0.2	0.2	0.5	1.0	-
5	0.5	0.5	0.5	0.5	0.5	1.0	-
6	0.6	0.6	0.6	0.4	0.6	1.0	-
10	0.6	0.6	0.6	0.4	0.6	1.0	1.0
10*	0.3	0.3	0.3	0.2	0.3	0.5	0.5

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 documents 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 Run1, there are respectively 21/9/16 topics for which R-Precision is better/worse/tied. For Run10, the corresponding figures are 23/9/14 topics.

Let us look at the results in more detail. Run10 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.

	HARD .			**SOFT**		
RUN	**Avg Prec.**	**Rels@10**	**R-Prec.**	**Avg Prec.**	**Rels@10**	**R-Prec.**
Baseline	*0.259* *(0.251)*	*3.11* *(3.23)*	*0.250* *(0.246)*	*0.257* *(0.229)*	*3.84* *(3.32)*	*0.275* *(0.213)*
1	0.251 (0.240) -3.02%	3.24 (3.39) 4.18%	**0.262 -/**** **(0.245)** **4.85%**	0.241 (0.222) -6.04%	3.76 (3.51) -2.08%	0.25 (0.222) -7.7%
5	0.2432 (0.242) -5.92%	3.20 (3.33) 2.89%	0.261 (0.249) 4.53%	0.235 (0.224) -8.26%	3.69 (3.50) -3.91%	0.248 (0.223) -9.84%
6	0.242 (0.238) -6.46%	3.11 (3.35) 0%	0.251 (0.243) 0.44%	0.231 (0.220) -10.21%	3.58 (3.49) -6.77%	0.246 (0.221) -10.64%
*10/10**	**0.258 -/**** **(0.240)** **-0.35%**	3.44 (3.39) 10.61%	**0.2801 */**** **(0.250)** **12.17%**	0.246 (0.219) -4.09%	3.89 (3.50) 1.3%	0.2567 (0.220) -6.76%

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 Run1 and Run10 to particular sources of evidence.

5.3 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.

Using the *familiarity/much* source improves the baseline significantly, but using the *familiarity/little* source does not. It would seem that our new language modelling approach to familiarity ranking is highly effective for high familiarity topics, and this is partly due to the relatively high baseline performance for these topics. Users, who are highly familiar with a topic, generate more effective topic specifications. For the *familiarity /much* subset, the baseline R-Precision is 0.307 compared with 0.250 over all topics. Consequently, it is likely that better *familiarity/much* queries will be generated given that we assume the top 10 documents in the baseline are pseudo-relevant.

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, 0.337) 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 is likely that the very different retrieval mechanisms ranked the documents differently, and that the combination resulted in improved ranking. This phenomenon has been observed generally in fusion/evidence combination.

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.

Second source	# topics	Base	Second source weights and R-Precision					
		R-Pr	wgt	R-Pr	wgt	R-Pr	wgt	R-Pr
Genre/ oped	7	0.165	0.5	0.198	0.25	**0.251**	0.125	0.189
Geog/ US	15	0.190	0.5	0.176	0.25	0.176	0.125	**0.193**
Geog/ Non-US	7	**0.337**	0.5	0.190	0.25	0.254	0.125	0.257
Fam/ little	27	0.208	0.5	0.210	0.25	0.200	0.125	0.197
Fam/ much	19	0.307	0.5	0.395 -/**	0.25	**0.393** */**	0.125	0.357 */**
Rerank baseline	45	0.250	1.0	**0.326** */**	0.75	0.292 */**	0.5	0.268

If you compare the 'Rerank baseline' run in Table 4, with any of the evidence combination runs in Table 3, you will note that this simple combination of baseline and re-ranked baseline achieves an R-Precision of 0.326 over all topics, and outperforms the best of the evidence combination runs. That is, a purely topicality based combination, achieves better hard R-Precision, than combinations of techniques specifically devised to improve hard effectiveness. The re-ranking of the baseline is performed using the Lemur KL method that implements a form of pseudo-relevance feedback based on the Kullback-Leibler divergence. And, our familiarity ranking methods are similarly based. We therefore decided to run a further set of experiments comparing the effectiveness of the KL-representative queries, the KL-discriminative queries, and the LEMUR KL method.

We generated KL-discrimination queries for **all** topics, and re-ranked the baseline resulting in the DQ (discriminating queries) run. Similarly, we generated KL-representation queries for all topics, and re-ranked the baseline resulting in the RQ (representative queries) run. We then combined the baseline with the DQ run (weighting 1.0 and 0.25), and the baseline with the RQ run (weighting 1.0 and 0.5). The weightings were selected to given the best overall result. Given these combination runs are the equivalent to pseudo-relevance feedback runs, we compared them against the re-ranking of the baseline using the Lemur/KL method, which itself is a pseudo-relevance feedback method. The results of these runs are given in Table 5. We report R-Precision for the set of all (ALL) topics, and for the *familiarity/much* (MUCH) and *familiarity/little* (LITTLE) subsets.

Table 4: Document-level evaluations of pseudo-relevance feedback runs. R-Precision using hard and soft relevance judgments. */** indicate statistical significance compared with baseline at level 0.05 using Wilcoxon/Sign test respectively, for HARD results only.

	Base		Base + DQ		Base + RQ		KL (reranked baseline)	
	HARD	*SOFT*	*HARD*	*SOFT*	*HARD*	*SOFT*	*HARD*	*SOFT*
ALL	**0.249**	0.273	**0.297** */**	0.295	**0.265** */**	0.312	**0.311** */**	0.318
MUCH	**0.307**	0.356	**0.393** */**	0.375	**0.341** -/**	0.384	**0.394**	0.399
LITTLE	**0.206**	0.214	**0.227**	0.236	**0.210**	0.259	**0.250**	0.258

Overall, the performance of the pseudo-relevance feedback runs is significantly better than the baseline run, and the KL run is marginally better than the Base/DQ run.

For the MUCH topic subset, Base/DQ and KL are comparable. Indeed, Base/DQ is significantly better than the baseline, whereas KL is not. Examining the soft effectiveness, it appears that KL achieves improvements in hard effectiveness through corresponding improvements in soft effectiveness. Base/DQ achieves comparable hard effectiveness with comparatively lower levels of soft effectiveness. This provides some evidence that the discriminating queries (DQ), when used for the *familiarity/much* topics, do in fact boost **hard** effectiveness as measured by R-Precision. Interestingly, the representative queries (RQ) also achieve good levels of performance for this topic subset. As observed earlier, the MUCH topics provide a better starting point for any pseudo-relevance process, given their superior baseline performance.

Neither the DQ nor RQ queries perform significantly better than the baseline for the *familiarity/little* topics. Interestingly, the KL run is best for these topics.

It would seem there is evidence that combining the KL run with the baseline is likely to result in improvements for all topics, as demonstrated in Table 3. The discriminating queries are highly effective when applied to the *familiarity/much* topics, and it will be interesting to see how to combine the baseline, the KL re-ranking and the DQ/much (DQ for much topics only) sources of evidence to achieve optimal performance.

6. Conclusions

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

- Our new topic modelling approach to generating discriminating queries, based on the Kullback-Leibler divergence, is highly effective for the *familiarity/much* topics, and significantly so compared with the baseline. However, comparable levels of performance are achieved for the *familiarity/much* topics using the standard KL method implemented in the Lemur toolkit. But, there is some evidence that the discriminating queries are boosting hard performance rather than simply boosting soft performance.
- 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 queries 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.

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 our colleagues at RGU and Rutgers for their feedback on this work, and in particular Nick Belkin. We would like to thank Paul Ogilvie for his invaluable advice on using the Lemur toolkit, and Diane Kelly for her insightful feedback on this work.

References

Allan, J. HARD track overview in the TREC 2003 High Accuracy Retrieval from Documents. In: The Twelfth Text REtrieval Conference, TREC 2003. E.M. Voorhees & L.P. Buckland [eds.]. (pp. 24-37). GPO, Washington, D.C., 2004.

Allan, J. HARD track overview in the TREC 2004 (Notebook) High Accuracy Retrieval from Documents. In: The 13th Text REtrieval Conference, TREC 2004. GPO, Washington, D.C., 2004.

Aslandogan, Y.A and Yu, C.T. (2000) Multiple Evidence Combination in Image Retrieval: Diogenes Searches for People on the Web. In: Proceedings of SIGIR 2000, Athens, Greece.

Belkin, N.J., Kelly, D., Lee, H.-J., Li, Y.-L., Muresan, G., Tang, M.-C., Yuan, X.-J. & Zhang, X.-M. Rutgers' HARD and web interactive track experiences at TREC 2003. In E.M. Voorhees & L.P. Buckland (Eds.) The Twelfth Text REtrieval Conference, TREC 2003 (pp. 532-543). GPO, Washington, D.C., 2004

Croft, W.B. and Lafferty, J. (Eds.): Language modeling for information retrieval. Kluwer Academic Publishers, The Netherlands, 2003.

Jose, J.M. (1998) An integrated approach for multimedia information retrieval, PhD thesis, School of Computing and Mathematical Sciences, Robert Gordon University, Aberdeen, UK.

Jose, J.M. and Harper, D.J. (1997) A retrieval mechanism for semi-structured photographic collections, in *Proceedings of the DEXA'97 Conference*, number 1308 in Lecture Notes in Computer Science, pages 276-292, Springer.

Jose, J.M. and Harper, D.J. (1998) Spatial querying for image retrieval: A user-oriented evaluation, in Proceedings of SIGIR'98, pages 232-240, Melbourne, Australia.

Lafferty, J. & Zhai, C. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM-SIGIR conference on research and development in information retrieval*, W.B. Croft, D.J. Harper, D.H. Kraft, & J. Zobel (Eds.), New Orleans, Louisiana 2001 (pp.111-119).

Urban, J., Jose, J. M., Van Rijsbergen, C. J. (2003) An Adaptive Approach Towards Content-Based Image Retrieval, in Proceedings of the 3rd International workshop on Content Based Multimedia Indexing, pages 119-126, Rennes, France.