University of Glasgow at TREC 2012: Experiments with Terrier in Medical Records, Microblog, and Web Tracks

Nut Limsopatham, Richard McCreadie, M-Dyaa Albakour,
Craig Macdonald, Rodrygo L. T. Santos, and Iadh Ounis
{nutli,richardm,dyaa,craigm,rodrygo,ounis}@dcs.gla.ac.uk

School of Computing Science
University of Glasgow
Glasgow, UK

ABSTRACT

In TREC 2012, we focus on tackling the new challenges posed by the Medical, Microblog and Web tracks, using our Terrier Information Retrieval Platform. In particular, for the Medical track, we investigate how to exploit implicit knowledge within medical records, with the aim of better identifying those records from patients with specific medical conditions. For the Microblog track adhoc task, we investigate novel techniques to leverage documents hyperlinked from tweets to better estimate relevance of those tweets and increase recall. Meanwhile, for the Microblog track filtering task, we developed a new stream processing infrastructure for real-time adaptive filtering on top of the Storm framework. For the TREC Web track, we continue to build upon our learning-to-rank approaches and novel xQuAD framework within Terrier, increasing both effectiveness and efficiency when ranking.

1. INTRODUCTION

In TREC 2012, we participate in Web adhoc and diversity tasks, the Microblog real-time adhoc and filtering tasks, and the Medical Records track. Our focus is the development of effective and efficient approaches to these tasks within the Terrier Information Retrieval (IR) platform [24]. In line with the vision for Terrier [16], our participation revolves around the development of effective and efficient data-driven ranking models and new infrastructures for real-time streaming applications.

In the Medical Track, our aim is to develop effective approaches to identify medical records from patients with specific medical conditions. This task is challenging since medical records often contain implicit knowledge that can mislead an IR retrieval system if not placed in context. We propose three novel approaches to tackle medical record retrieval that exploit this implicit knowledge by incorporating term context, topic modelling and knowledge-based reasoning, respectively.

We participate in both the adhoc and filtering tasks of the Microblog track. The major goal of our participation in both tasks is to develop effective real-time approaches to rank tweets by their relevance and quality for a user query. For the adhoc task, we proposed three novel approaches that leverage the documents hyperlinked from the tweets to better estimate when they are relevant. Meanwhile, for the filtering task, we develop a new stream processing infrastructure for real-time adaptive filtering on top of the Storm framework.[1] Furthermore, in both tasks, we leverage data-driven learning to improve effectiveness.

In our participation in the Web track, our primary goal is to enhance our data-driven learning infrastructure and continue to improve our state-of-the-art xQuAD framework for search result diversification. In particular, for the adhoc ranking task, we deploy a state-of-the-art learning-to-rank algorithm in combination with xQuAD to diversify the sample of documents to be ranked. For the diversification task, we employ xQuAD to diversify a learned baseline and propose a new approach to learn how to diversify Web search results within xQuAD.

The remainder of this paper is structured as follows. In Section 2, we describe our participation in the Medical Records track. Section 3 and 4 detail our participation in the new Microblog track real-time adhoc and filtering tasks, respectively. In Sections 5 and 6, we describe our Web track adhoc task and Web track diversity task participations, respectively. Conclusions are provided in Section 7.

2. MEDICAL RECORDS TRACK

The aim of our participation in the TREC Medical Records track is to investigate novel voting-based approaches to tackle the challenge of implicit knowledge within medical records. Such knowledge is known among health practitioners but is hidden from search systems [9, 10, 13]. We detect this knowledge at different levels of the retrieval process (sentence, record and inter-record levels) and propose three approaches to do so, namely: Medical Context Integration; Medical Topic Modelling; and Medical Knowledge-Based Reasoning.

We build upon a common baseline approach that uses the successful Voting Model [14] to ranks patient visits as part of an extended voting process. In particular, we deploy the expCombSUM voting technique [14] to rank patient visits, accumulating the relevance scores of medical records [12] computed using the parameter-free DPH weighting model [2] with Bo1 query expansion [1]. Each of our three proposed approaches extend this baseline as follows.

Our first approach – Medical Context Integration – leverages implicit knowledge from within records at the sentence level. Building upon our effective negation handling approach [12, 13], we leverage the context (positive or negative, present or history, patients' condition or family mem-

[1] `https://github.com/nathanmarz/storm/wiki/Rationale`

bers' condition) of medical terms from each sentence to improve retrieval effectiveness. Our second approach – Medical Topic Modelling – accounts for the relationships between medical records (e.g. their issuing hospital department). In particular, we use the Latent Dirichlet Allocation (LDA) topic modelling technique [6] to extract latent topics from all medical records in the collection, using the MALLET toolkit [18]. We combine the retrieval score for each medical record with the similarity between the topics extracted from those records and the query. Finally, our third proposed approach – Medical Knowledge-Based Reasoning – exploits various domain-specific resources (e.g. controlled-vocabularies, ontologies, health-related websites) from within the medical domain to better represent the content of each medical record. In particular, we use MetaMap [4] to identify concepts related to four medical aspects[2] in each medical record and query [11]. Using the concepts identified from a query, we expand that query with other related concepts, e.g. the concept heart disease (diagnosed condition) is related to the concept amiodarone (a drug used to combat it). Each medical record is scored using a mixture model that favours records containing more of the (expanded) query's concepts.

We submitted 4 runs to the TREC 2012 Medical Records track. Each run uses the Terrier IR platform [24] to retrieve medical records. Any parameters (e.g. the number of latent topics for LDA) were trained using the 34 topics of the TREC 2011 Medical Records track. Note that these runs build upon our baseline voting approach that incorporates query expansion. Each run is described below:

- **uogTrMConQ**: deploys our Medical Context Integration approach.
- **uogTrMConQT**: uses our Medical Topic Modelling approach on top of uogTrMConQ.
- **uogTrMConQRa**: employs our Medical Knowledge-Based Reasoning approach on top of uogTrMConQ.
- **uogTrMConQRd**: uses the same Medical Knowledge-Based Reasoning approach as uogTrMConQRa, but uses fewer concepts (only those associated to the diagnosis and treatment aspects).

In addition, we report the performance of five unsubmitted runs as baselines, representing traditional retrieval models, retrieval with query expansion and the negation handling approach we proposed for the TREC 2011 Medical track:

- **uogTrMB**: applies only the DPH weighting model and the expCombSUM voting technique.
- **uogTrMNeg**: deploys our NegFlag representation [12], the DPH weighting model and the expCombSUM voting technique.
- **uogTrMCon**: uses the same techniques as uogTrMConQ, but the DFR Bo1 query expansion is not applied.
- **uogTrMBQ**: applies DFR Bo1 on uogTrMB.
- **uogTrMNegQ**: applies DFR Bo1 on uogTrMNegQ.

Run	Submitted	infAP	infNDCG	R-prec	P@10
TREC Median	N/A	0.1689	0.4244	0.2961	0.4702
uogTrMB	✗	0.1703	0.4167	0.3057	0.4638
uogTrMNeg	✗	0.1833	0.4355	0.3300	0.4894
uogTrMCon	✗	**0.1869**	**0.4412**	**0.3313**	**0.4936**
uogTrMBQ	✗	0.1871	0.4408	0.3197	0.4617
uogTrMNegQ	✗	0.2011	0.4567	0.3361	0.5170
uogTrMConQ	✔	0.2033	0.4575	0.3371	0.5213
uogTrMConQT	✔	0.1909	0.4454	0.3260	0.4957
uogTrMConQRa	✔	0.2258	0.5084	**0.3699**	0.5426
uogTrMConQRd	✔	**0.2313**	**0.5085**	0.3663	**0.5532**

Table 1: Results of our runs on TREC 2012 Medical Records track's official measures.

Table 1 shows the results of our submitted and additional runs comparing to the TREC median. First, we find that all of our runs provide similar or improved performance in comparison to the TREC Median under all reported measures. This shows that our approaches are effective for ranking patient visits. Next, comparing our Medical Context Integration approach (uogTrMConQ) to our five baselines, we observe that it outperforms all five baselines for all measures. This result suggests that it is important to effectively handle the contexts of terms when searching medical records. Our remaining three runs build upon uogTrMConQ, hence we treat it as a new baseline. Examining our Medical Topic Modelling approach (uogTrMConQT), we see that it provides similar performance to our new baseline uogTrMConQ, indicating that modelling the topics in medical records does not enhance retrieval effectiveness. Finally comparing the two approaches that use our Medical Knowledge-Based Reasoning approach (uogTrMConQRa and uogTrMConQRd), we observe that both outperform our new baseline (uogTrMConQ) by a large margin (e.g. 0.4575 to 0.5084 for uogTrMConQRa under infNDCG). Hence, we conclude that our knowledge-based reasoning technique is highly effective. Overall, we have shown that enhancing the representation of the query with inferenced medical concepts can improve the retrieval effectiveness.

3. MICROBLOG TRACK: REAL-TIME ADHOC TASK

The aim of the TREC Microblog track real-time adhoc task is to retrieve both recent and relevant tweets to a user query at a point in time. For our participation in the adhoc task, we investigate novel approaches that leverage the documents hyperlinked from the tweets to better estimate how relevant each tweet is to the query. Our aim is to find those tweets that are relevant to the query because they link to a good quality document, rather than relying only on the content of the tweet. In particular, we propose three unique approaches, each of which leverage the documents hyperlinked from a tweet in a different manner, namely Basic Sample Expansion (BsE); Learned Sample Expansion (LsE); and Cross-Index Document Expansion (CIDE).

The Tweets2011 corpus is dissimilar to traditional TREC datasets, in that it is not directly provided by TREC. Instead, a set of approximately 16 million tweet identifiers are provided along with a tool for downloading the tweets for those tweet identifiers [25]. The corpus is dynamic, i.e. the number of available tweets changes over time, as users delete

[2] symptom, diagnostic test, diagnosis, and treatment [37]

Data	Quality	2011	2012
Tweets2011 Corpus	Time Range	23/01/11 → 08/02/11	
	Days	16	16
	# Tweets	15,663,909	10,561,763
	Avg. Tweets Per Day	978,994	660,110

Table 2: Tweets2011 corpus statistics.

Approach	Submitted	Uses Hyperlinked Documents	R-Precision	P@30	Recall
TREC Median	N/A	N/A	—	0.0687	—
DPH+QE	✗	✗	**0.2225**	0.1969	0.7294
uogTrBsE	✓	✓	0.2188	0.2062	0.7344 ▲ ▲
uogTrLsE	✓	✓	0.2208	**0.2232** ▲	**0.7508** ▲ ▲
uogTrCIDE	✓	✓	0.1059	0.1124	0.6687

Table 3: High-Rel performance of our TREC 2012 Microblog track adhoc task runs in comparison to the TREC median. The best performing run under each measure is highlighted in bold. Statistical significance (paired t-test) at p<0.05 over the baseline is denoted ▲, while statistical significance (paired t-test) at p<0.01 over the baseline is denoted ▲ ▲.

their own tweets, or spam accounts are removed by Twitter itself [36]. We originally downloaded the Tweets2011 corpus for TREC 2011 using a customised version of the tool that scrapes the HTML page for each tweet to obtain its content. Although a study showed that the increasing unavailability of some tweets over time did not effect the ranking of systems submitted to TREC 2011 [21], to assure that each group had access to as similar a set of tweets as possible, TREC provided a list of tweetIDs to be used by participants for the 2012 task. We filtered our original copy of the Tweets2011 corpus using this tweetID list. Table 2 lists the statistics of both our copies of the Tweets2011 corpus. The 2011 column provides the statistics of our original version, while the 2012 column lists the statistics of the filtered version used for TREC 2012. We also downloaded the pages hyperlinked from each tweet, forming an aligned corpus. Both of these corpora were indexed[3] using the Terrier IR platform [24], removing stopwords and stemming each tweet using the English Porter Stemmer.

To generate our runs, we first create a time-ordered ranking of tweets from Tweets2011 for each topic. In particular, for a topic we retrieve a ranking of tweets containing one or more query terms in a reverse-chronological order that were posted before the timestamp. For ranking we use the DFReeKLIM weighting model [3] and apply query expansion. In this case, each query is expanded using the top 20 terms from the top 20 tweets retrieved using DFReeKLIM. Each expansion term is scored using the Kullback-Leibler (KL) divergence measure [1]. We use this ranking as our baseline for the adhoc task, denoted *DPH+QE*. For each query, we also rank the documents that were hyperlinked from a tweet posted before the query timestamp, using the parameter-free DPH weighting model.

We submitted three runs to the TREC 2011 Microblog track, each of which uses an novel approach to improve ranking performance over the baseline using the hyperlinked documents from each tweet. In particular, our proposed Basic Sample Expansion (BsE) approach uses a heuristic to combine the DFReeKLIM score for the tweet and the DPH score for any document hyperlinked from that tweet. Here, we sum the two scores, where the tweet score receives 90% weight, while the hyperlinked document score receives 10% weight. Our proposed Learned Sample Expansion (LsE) approach similarly combines the scores for the tweet and any hyperlinked documents. In this case, we use the AFS learning-to-rank technique [22] to learn an effective combination of the tweet and hyperlinked document scores. This approach also uses an additional 46 features describing the tweet and recent discussions from Twitter relating to the query. To train this approach, we used the TREC 2011 Microblog track topics. Finally, our Cross Index Document Expansion (CIDE) approach uses a different method for leveraging the hyperlinked documents. In particular, under this approach, we build a virtual document for tweets that have at least one hyperlinked document, comprised of both the terms within each tweet and the terms within any hyperlinked documents. The DFReeKLIM score of the virtual document is then used to rank tweets.

Our three runs are as follows:

- **uogTrBsE**: DPH+QE with Basic Sample Expansion.
- **uogTrLsE**: DPH+QE with Learned Sample Expansion.
- **uogTrCIDE**: DPH+QE with Cross Index Document Expansion.

Table 3 reports the performance of our three runs in terms of R-Precision, P@30 and Recall. The best performing run under each measure is highlighted in bold. Statistical significance using the paired t-test in comparison to the baseline at confidences $p < 0.05$ and $p < 0.01$ are denoted with ▲ and ▲ ▲, respectively. From Table 3, we observe the following. First, our uogTrBsE and uogTrLsE runs outperform the TREC median, indicating that they are effective approaches. Second, our baseline approach also outperforms that TREC median, indicating that it is a strong baseline. Third, uogTrCIDE achieved a lower performance than both the TREC Median and our baseline, indicating that combining the tweets and hyperlinked documents into larger virtual documents is not effective for the 2012 topics. Fourth, both our uogTrBsE and uogTrLsE approaches increase recall by a statistically significant margin (paired t-test $p < 0.01$), showing how hyperlinked documents can be leveraged to introduce more relevant tweets into the final ranking. Finally, our uogTrLsE approach outperforms the baseline under P@30 by a significant margin (paired t-test $p < 0.05$), indicating that using learning-to-rank is able to increase ranking effectiveness in the top ranks. Figure 1 illustrates the ROC curve for each of our three runs, in addition to the baseline. From Figure 1, we see that uogTrBsE and uogTrLsE are more effective than either the baseline or uogTrCIDE. Overall, we have shown that by leveraging the documents hyperlinked from tweets, retrieval effectiveness can be significantly improved.

4. MICROBLOG TRACK: REAL-TIME FILTERING TASK

The primary aim of our participation in the real-time filtering task is to extend our Terrier IR platform to support

[3] Tweet indexing under Terrier requires an additional plugin available from `http://ir.dcs.gla.ac.uk/wiki/Terrier/Tweets11`

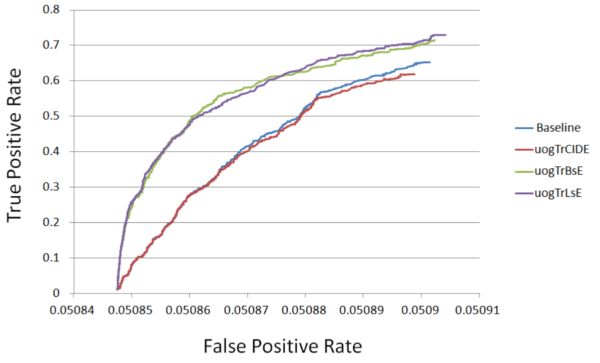

Figure 1: ROC Curve for our submitted runs in comparison to a baseline without any sample expansion technique applied.

real-time filtering. In particular, we develop a novel stream processing infrastructure based on the open source Storm framework. We use this infrastructure to perform adaptive filtering by building and adapting a profile of the user's interests over time. This profile is consulted for each incoming tweet to make a decision whether it is relevant to the user or not.

Adaptive filtering was previously investigated in the TREC Filtering track [28] within the news domain. The main difference here is that the documents (tweets) are much shorter than news articles. Also, the starting user profile is small and only consists of a query and a single tweet. We approach the problem of adaptive filtering by employing a common method used extensively in the TREC Filtering track [28]. This method is based on a classifier that uses the popular Rocchio's relevance feedback approach [29] to build a profile of the user's interests and update it online using the explicit judgements provided by the user. More specifically, at each point of time t the profile of the user is represented by a vector $\vec{c}_t$ which is called the *centroid* of the user's interests. The centroid is calculated using Equation (1):

$$\vec{c}_t = \frac{1}{|R_t|} \cdot \sum_{d_i \in R_t} \vec{d}_i \tag{1}$$

where R_t is the set of tweets judged relevant by the user so far (at time t). For each incoming tweet, the cosine similarity is computed between the centroid and the tweet. When the cosine value exceeds a certain threshold θ, the tweet is retrieved, otherwise it is not. The centroid is initially represented with a single tweet, which is a short document. To tackle this sparsity problem, we apply query expansion using the method in [1] to derive terms for a richer representation of the centroid. For each topic, query expansion is applied using the query and an index of all the tweets in the corpus prior to the first relevant tweet. Moreover, we investigate using the top retrieved tweets from that index to enrich the representation of the centroid. Finally, we also explore a method to vary the threshold θ according to the values of query independent features. Using the training topics, both the occurrence of a URL and the occurrence of a hashtag in a tweet are identified as good features that indicate relevance. Following this, we use simple heuristics to select different values for the threshold when the tweet contains URLs and/or hashtags.

In our experiments, we use language modelling with Dirichlet prior [39] for weighting the terms of the tweets in the vector representation. Tweets that do not contain at least one query term are not considered for similarity computation and are regarded as irrelevant. We also experiment with relaxing this condition by considering tweets which contain at least one term in either the query or the first tweet. The threshold θ is optimised for both filtering measures of the TREC Filtering track 2002 (F_0.5 and T11SU) [28] using the ten training topics provided by TREC. With this experimental setup, we submitted four different runs to the real-time filtering task:

- **uogTrFFDmN**: The centroid is initialised with the starting tweet and the query. Tweets that do not contain at least one query term are considered irrelevant and no similarity computation is performed.

- **uogTrFFeDm**: The centroid is enriched with terms derived with query expansion. The condition for performing the similarity computation is the one used in the previous run uogTrFFDmN.

- **uogTrFADmN**: Terms derived with query expansion and top retrieved tweets are used to enrich the initialisation of the centroid. Tweets that do not contain at least one query term or one term from the first tweet are considered irrelevant and no similarity computation is performed.

- **uogTrFADmI**: The threshold is varied according to the occurrence of the query independent features (URL and hashtags). The condition for performing the similarity computation is the one used in the previous run uogTrFADmN.

	set_prec	set_recl	F_0.5	T11SU
TREC Median	0.1767	0.3343	0.1491	0.2076
uogTrFFDmN	**0.5508**	0.1394	0.1915	0.3427
uogTrFFeDm	0.4940	0.1621	0.2095	0.3300
uogTrFADmN	0.4206	0.3370	**0.3435**▲ ▲	**0.3615**△ △
uogTrFADmI	0.3197	**0.3868**	0.2537	0.2510

Table 4: Results of the submitted runs to the filtering task of the Microblog track. ▲ ▲ denotes statistically significant improvement over all other runs. △ △ denotes statistical significant improvement over uogTrFADmI. Statistical significance is estimated with a paired t-test at $p < 0.01$. Figures in bold represent the top performances.

The performance of the submitted runs is reported in Table 4 in terms of set precision, set recall, F measure with $\beta = 0.5$, and the utility measure T11SU. We summarise the findings as follows. All our runs outperform the TREC median using either of the official measures (F_0.5, and T11SU). A common baseline approach in an adaptive filtering task that simply retrieves no documents (tweets) would achieve a utility T11SU of 0.33. All our runs, apart from uogTrFADmI, improve on this baseline as well. The best performing run in terms of the two official measures is uogTrFADmN, which shows the power of initialising the centroid of the classifier with an enriched representation using query expansion and top retrieved tweets. Finally, using the query independent features does not improve the performance. However, it achieves a higher recall than the recall obtained with any other run at the cost of harming the precision. As a summary, we have developed a novel Storm-based infrastructure suitable for real-time adaptive filtering and we have shown that enriching the initial user's profile leads to significantly more effective filtering performance over time.

5. WEB TRACK: ADHOC TASK

In the adhoc task, our primary aim is to enhance our data-driven learning infrastructure that has proven effective during previous participations [19, 33], such that it becomes easier to generate and evaluate many learning-to-rank models in a short timeframe. To this end, we deploy Terrier using a framework for the fast computation of document features to use with learning-to-rank, as discussed in [20]. Then, we enhance this framework by deploying a state-of-the-art learning-to-rank algorithm based on gradient-boosted regression trees [38]. Thereafter, our investigation encompasses three aims. Firstly, we aim to improve the recall of the sample of documents to be re-ranked by the learned model. In particular, we introduce documents that would have ranked highly for common reformulations of the user's initial query, but not for the initial query. Secondly, we examine how these common reformulations can be used to improve the precision of the top-ranked results. Lastly, we propose a novel approach that measures the contribution of groups of documents from the same domain as learning features.

We index the category A (∼500M English documents) and category B (∼50M English documents) subsets of the ClueWeb09 corpus without stemming or stopwords. At retrieval time, a weak Porter stemmer and the DPH [2] weighting model are used to identify 5000 documents to re-rank using the learned models. This setting has been shown effective for ClueWeb09 [17]. Our category A and B runs both use a total of 44 features, as described in Table 5. For learning and training hyper-parameters, we employ the state-of-the-art LambdaMART learning-to-rank technique [8, 38],[4] with the 150 queries from the TREC Web tracks 2009-2011 used as training data and validation data. The 150 queries are randomly split into training (60%) and validation (40%), so as to prevent overfitting.

We submitted three runs to the adhoc task, in addition to two submitted baselines runs:

- **uogTrA44** (unsubmitted): 44 features for the category A corpus, learned using LambdaMART.
- **uogTrA44s9**: As uogTrA44, but encapsulates more relevant documents in the sample from related sub-queries (as identified in Section 6), with a weight of 90% to the original query, and 10% to the related sub-queries.
- **uogTrA44xi**: As uogTrA44, but additionally applies xQuAD (Section 6) with uniform sub-query probabilities over the related sub-queries.
- **uogTrB44** (unsubmitted): 44 features for the category B corpus, learned using LambdaMART.
- **uogTrB45aIs**: The 44 features from uogTrB44 plus one additional feature that measures, for a given document, the contribution of groups of documents from the same domain to the relevance of the document.

Run	Submitted	Category	NDCG@20	ERR@20
TREC median		A	0.1023	0.1436
uogTrA44	✘	A	0.1938	0.2644
uogTrA44s9	✔	A	0.1938	0.2644
uogTrA44xi	✔	A	0.2383	0.3132
TREC median		B	0.1243	0.1872
uogTrB44	✘	B	0.1852	0.2592
uogTrB45aIs	✔	B	0.1831	0.2554

Table 6: Results of the three submitted runs for the Web track adhoc task under the normalised discounted cumulative gain at rank 20 (NDCG@20) and expected reciprocal rank at rank 20 (ERR@20) measures.

Table 6 reports the performance of the three submitted adhoc runs under the normalised discounted cumulative gain at rank 20 (NDCG@20) and expected reciprocal rank at rank 20 (ERR@20) measures. In particular, we observe firstly that all our submitted runs are markedly above the median performance of their respective categories. Next, our method to improve the sample (uogTrA44s9) does not impact upon the ranking of the top 20 documents compared to the baseline uogTrA44. In contrast, uogTrA44xi, which applies xQuAD to the baseline, results in a marked increase in both NDCG@20 and ERR@20. Indeed, uogTrA44xi was the best TREC Web track adhoc task run. For the B category, we find that our feature that measures the relevance of documents from the same domain actually decreases performance. Overall, we conclude that while the xQuAD framework is designed for the diversification of search results, it can also markedly improve the adhoc effectiveness of a state-of-the-art learning-to-rank technique.

6. WEB TRACK: DIVERSITY TASK

In the diversity task, we continue to improve our state-of-the-art xQuAD framework for search result diversification [30, 31, 32, 35]. In particular, xQuAD models an ambiguous query as a mixture of the multiple possible information needs underlying this query, represented as different *sub-queries* [34]. Given an initial ranking $\mathcal{R}$ for the query q, and a set of sub-queries $\mathcal{S}$ identified for this query, xQuAD iteratively builds a re-ranking $\mathcal{D}$ by selecting, at each iteration, a document $d^* \in \mathcal{R} \setminus \mathcal{D}$ such that:

$$d^* = \mathop{\arg\max}_{d \in \mathcal{R} \setminus \mathcal{D}} \; (1-\lambda)\, p(d|q) + \lambda\, p(d, \bar{\mathcal{D}}|q), \tag{2}$$

where $p(d|q)$ and $p(d, \bar{\mathcal{D}}|q)$ denote the *relevance* and the *diversity* of the document d in light of the query q. The latter can be interpreted as the probability that d, but none of the documents in $\mathcal{D}$, selected in previous iterations, satisfies the multiple sub-queries of q. In particular, the probability $p(d, \bar{\mathcal{D}}|q)$ can be further expanded as follows:

$$p(d, \bar{\mathcal{D}}|q) = \sum_{s \in \mathcal{S}} p(s|q)\, p(d|q,s) \prod_{d_j \in \mathcal{D}} p(\bar{d_j}|q,s), \tag{3}$$

where $s \in \mathcal{S}$ is one of the possible sub-queries underlying the query q, $p(s|q)$ conveys the importance of this sub-query in

[4] `http://code.google.com/p/jforests/`

Features	Total
Weighting models (DPH [2], PL2 [2], BM25 [27], LM, MQT [32])	21
Fields-based models (PL2F [15])	1
URL and link analysis features (e.g. PageRank)	11
Quality features (e.g. fraction of stopwords, table text [5])	8
Spam feature (Cormack's fusion score [7])	1
Term-dependence models (MRF [23], pBiL [26])	2
TOTAL	44

Table 5: Document features used in the Web track, both Category A and Category B runs.

light of q, $p(d|q,s)$ estimates the coverage of d with respect to s, and $\prod p(\bar{d}_j|q,s)$ estimates the novelty of any document satisfying s, given how badly this sub-query is satisfied by the previously selected documents $d_j \in \mathcal{D}$. In order to produce refined estimations of the relevance of a document to the initial query (i.e. $p(d|q)$), and the coverage of this document with respect to each of the sub-queries identified for this query (i.e. $p(d|q,s)$), we leverage the ranking models learned for our adhoc runs in Section 5. The probabilities of relevance and diversity are mixed through the diversification tradeoff λ. Likewise, the total number of iterations performed—i.e. the total number of documents from the initial ranking $\mathcal{R}$ to be diversified—is determined by another parameter, namely, the diversification cutoff τ. Both parameters λ and τ are optimised on training data.

In addition to xQuAD, we experiment with a novel machine learning approach for diversifying the search results. In particular, we propose a meta-learning approach that directly leverages existing learning-to-rank algorithms in order to produce effective ranking models that reward diversity. To this end, our approach extends a standard feature space for learning-to-rank (including, for instance, query-dependent and query-independent document features) into a space augmented towards two orthogonal axes. On the *coverage* axis, the feature space is augmented to include features that estimate the relevance of a document to multiple sub-queries. On the *novelty* axis, the space is augmented by taking into account the diminishing relevance of each document in light of the selection of other documents to compose a diverse ranking. To instantiate our proposed approach, we leverage the same basic features used for our adhoc runs in Section 5. In addition, we produce *sub-query-dependent* features by recomputing all query-dependent features in Table 5 at the sub-query level, and then reaggregating at the query level. Precisely, under this formulation, a sub-query-dependent feature f can be computed according to:

$$p_f(d|q) = \Psi_{s \in S}\, p(s|q)\, p_f(d|q,s) \prod_{d_j \in S} p_f(\bar{d}_j|q,s), \qquad (4)$$

where Ψ is an aggregation function, such as summation.

We submitted three runs to the diversity task:

- Runs **uogTrA44xu** (cat. A) and **uogTrB44xu** (cat. B) deploy xQuAD with each document's coverage of different sub-queries estimated through learning-to-rank, and with both the diversification cutoff τ and the trade-off λ optimised using previous years' data.

- Run **uogTrA44xl** (cat. A) deploys our novel learning approach, which incorporates query-dependent, query-independent, and sub-query-dependent features in a single learned model to reward diversity.

Table 7 shows the diversification performance of our submitted runs to the diversity task, as well as their corresponding (unofficial) baseline adhoc runs. From the table, we first observe that our runs based on xQuAD, i.e., uogTrA44xu (cat. A) and uogTrB44xu (cat. B), substantially outperform both the TREC median as well as our adhoc baselines (uogTrA44 and uogTrB44, respectively). Indeed, uogTrA44xu was the overall best performing run in the diversity task of the TREC 2012 Web track. Regarding our novel learning approach, run uogTrA44xl (cat. A) also substantially outperforms the TREC median. On the other hand, it underperforms slightly compared to the uogTrA44 adhoc baseline, which suggests that the learned model requires more data to generalise well from training to test.

	Category	ERR-IA @20	α-nDCG @20	task
TREC median	A	0.3240	0.4370	
uogTrA44	A	0.4173	0.5345	✗ adhoc
uogTrA44xu	A	**0.5048**	**0.6061**	✔ diversity
uogTrA44xl	A	0.4009	0.5161	✔ diversity
TREC median	B	0.3639	0.4762	
uogTrB44	B	0.4324	0.5395	✗ adhoc
uogTrB44xu	B	**0.4785**	**0.5795**	✔ diversity

Table 7: Results of the submitted runs to the diversity task of the Web track.

7. CONCLUSIONS

In TREC 2012, we participated in Web adhoc and diversity tasks, the Microblog real-time adhoc and filtering tasks, and the Medical Records track using our Terrier IR platform. In particular, for the Web track, we focused on enhancing our data-driven learning infrastructure and continued to improve our state-of-the-art xQuAD framework for search result diversification. For the Microblog track, we proposed three novel approaches that leverage the documents hyperlinked from the tweets to improve adhoc tweet search and developed a new stream processing infrastructure for real-time adaptive filtering on top of the Storm framework. Finally, for the Medical Records track, we proposed three new approaches to exploit implicit knowledge within medical records in the form of term context, topic modelling and knowledge-based reasoning, respectively.

8. ACKNOWLEDGEMENTS

This work was funded as part of the EPSRC CROSS project, grant number EP/J020664/1 and the EU FP7 SMART project, grant number 287583.

9. REFERENCES

[1] G. Amati. *Probability models for information retrieval based on Divergence From Randomness.* PhD thesis, Univ. of Glasgow, 2003.

[2] G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi. FUB, IASI-CNR and Univ. of Tor Vergata at TREC 2007 Blog track. In *Proc. of TREC*, 2007.

[3] G. Amati, G. Amodeo, M. Bianchi, G. Marcone, C. Gaibisso, A. Celi, C. De Nicola and M. Flammini. FUB, IASI-CNR, UNIVAQ at TREC 2011. In *Proc. of TREC*, 2011.

[4] A. R. Aronson and F. Lang An overview of MetaMap: historical perspective and recent advances. *JAMIA*, 2010.

[5] M. Bendersky, W. B. Croft, and Y. Diao. Quality-biased ranking of web documents. In *Proc. of WSDM*, 2011.

[6] D. Blei, A. Ng, and M. Jordan. Latent dirichlet Allocation. *the Journal of Machine Learning Research*, 2003.

[7] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large Web datasets. *Inf. Retr.*, 2011.

[8] Y. Ganjisaffar, R. Caruana, C. Lopes. Bagging gradient-boosted trees for high precision, low variance ranking models. In *Proc. of SIGIR*, 2011

[9] N. Limsopatham, C. Macdonald, R. McCreadie and I. Ounis. Exploiting Term Dependence while Handling Negation in Medical Search. In *Proc. of SIGIR*, 2012.

[10] N. Limsopatham, C. Macdonald and I. Ounis. Aggregating Evidence from Hospital Departments to Improve Medical Records Search. In *Proc. of ECIR*, 2013.

[11] N. Limsopatham, C. Macdonald and I. Ounis. A Task-Specific Query and Document Representation for Medical Records Search. In *Proc. of ECIR*, 2013.

[12] N. Limsopatham, C. Macdonald, I. Ounis, G. McDonald, and M. M. Bouamrane. University of Glasgow at Medical Records Track 2011: Experiments with Terrier. In *Proc. of TREC*, 2011.

[13] N. Limsopatham, R. L. T. Santos, C. Macdonald, and I. Ounis. Disambiguating biomedical acronyms using EMIM. In *Proc. of SIGIR*, 2011.

[14] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *Proc. of CIKM*, 2006.

[15] C. Macdonald, V. Plachouras, B. He, C. Lioma, and I. Ounis. Univ. of Glasgow at WebCLEF 2005: Experiments in per-field normlisation and language specific stemming. In *Proc. of CLEF*, 2005.

[16] C. Macdonald, R. McCreadie, R. L. T. Santos and I. Ounis. From Puppy to Maturity: Experiences in Developing Terrier. In *Proc. of OSIR*, 2012.

[17] C. Macdonald, R. L. T. Santos, and I. Ounis. The Whens and Hows of Learning to Rank. *Information Retrieval*, 2012.

[18] A. K. McCallum. MALLET: A Machine Learning for Language Toolkit. `http://mallet.cs.umass.edu`, 2002.

[19] R. McCreadie, C. Macdonald, I. Ounis, J. Peng, and R. L. T. Santos. Univ. of Glasgow at TREC 2009: Experiments with Terrier—Blog, Entity, Million Query, Relevance Feedback, and Web tracks. In *Proc. of TREC*, 2009.

[20] R. McCreadie, C. Macdonald, R. L. T. Santos and I. Ounis. University of Glasgow at TREC 2011: Experiments with Terrier in Crowdsourcing, Microblog, and Web Tracks. In *Proc. of TREC*, 2011.

[21] R. McCreadie, I. Soboroff, J. Lin, C. Macdonald, I. Ounis, D. McCullough. On Building a Reusable Twitter Corpus. In *Proc. of SIGIR*, 2012.

[22] D. Metzler. Automatic feature selection in the Markov random field model for information retrieval. In *Proc. of CIKM*, 2007.

[23] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *Proc. of SIGIR*, 2005.

[24] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A high performance and scalable information retrieval platform. In *Proc. of OSIR at SIGIR*, 2006.

[25] I. Ounis, C. Macdonald, J. Lin, and I. Soboroff. Overview of the TREC-2011 Microblog Track. In *Proc. of TREC*, 2011.

[26] J. Peng, C. Macdonald, B. He, V. Plachouras, and I. Ounis. Incorporating term dependency in the DFR framework. In *Proc. of SIGIR*, 2007.

[27] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proc. of TREC*, 1994.

[28] S. E. Robertson and I. Soboroff The TREC 2002 Filtering track report. In *Proc. of TREC*, 2002.

[29] J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System*, 1971.

[30] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for Web search result diversification. In *Proc. of WWW*, 2010.

[31] R. L. T. Santos, C. Macdonald, and I. Ounis. Selectively diversifying Web search results. In *Proc. of CIKM*, 2010.

[32] R. L. T. Santos, C. Macdonald, and I. Ounis. Intent-aware search result diversification. In *Proc. of SIGIR*, 2011.

[33] R. L. T. Santos, R. McCreadie, C. Macdonald, and I. Ounis. University of Glasgow at TREC 2010: Experiments with Terrier in Blog and Web tracks. In *Proc. of TREC*, 2010.

[34] R. L. T. Santos and I. Ounis. Diversifying for multiple information needs. In *Proc. of DDR at ECIR*, 2011.

[35] R. L. T. Santos, J. Peng, C. Macdonald, and I. Ounis. Explicit search result diversification through sub-queries. In *Proc. of ECIR*, 2010.

[36] I. Soboroff, D. McCullough, J. Lin, C. Macdonald, I. Ounis and R. McCreadie. Evaluating Real-Time Search over Tweets. In *Proc. of ICWSM*, 2012.

[37] E. Silfen. Documentation and coding of ED patient encounters: an evaluation of the accuracy of an electronic medical record. *The American Journal of Emergency Medicine*, 2006.

[38] Q. Wu, C. J. C. Burges, K. M. Svore, J. Gao. Ranking, boosting, and model adaptation. Technical Report MSR-TR-2008-109, Microsoft, 2008.

[39] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 2004.