University of Wolverhampton at the TREC-2011 Microblog Track

Georgios Paltoglou and Mike Thelwall
University of Wolverhampton
Wulfruna Street, Wolverhampton
WV1 1LY, UK
{g.paltoglou,m.thelwall}@wlv.ac.uk

January 23, 2012

Abstract

In this report we discuss the experiments we conducted at the University of Wolverhampton for the Microblog Track at TREC-2011. As this was the first time we participated in TREC and the particular task presents some unique challenges we initially focused on properly analyzing and indexing the new Tweets2011 Corpus. We experimented with the effects that some standard IR techniques, such as query expansion and proximity models have in this setting. Initial results indicated that both techniques provide small increases in precision, but more experiments are needed for final conclusions to be reached. Lastly, we experimented with using the page that a tweet links to as part of the tweet. The results were particularly low, indicating a potential error in the indexing process or a natural outcome, due to the increased length of the combined documents. More research into answering the issue is underway.

1 Overview

This is first time that the University of Wolverhampton has participated in TREC. Due to time and resource constraints, we decided to use one of the publicly available IR toolkits, instead of building our own. Consequently, our primary goal for this year's participation was to decide which toolkit to use, get familiar with the functionalities that it offers, its capabilities and limitations, test some baseline retrieval algorithms and prepare for next year's TREC, where we could apply this year's knowledge to extend the chosen toolkit.

After a careful consideration [4] of the state-of-the-art, publicly available toolkits (e.g., Indri[1], Lemur[2], Lucene[3], Terrier[4]), we finally resorted to using the Terrier toolkit [5]. The toolkit's detailed documentation, significant extensibility and vibrant and helpful discussion forums were considered as important factors in the final decision.

This year we decided to only participate in the newly introduced Microblog track. The particular requirements of the track required that we change the way that Terrier ranks documents from the standard relevancy-based retrieval, to a time-dependable and time-constraint retrieval where tweets are ranked in order of posting time (with newer and relevant tweets first) and tweets which are published after the time of the query aren't considered for the estimation of global statistics, such as term IDF and average document length values.

We addressed the issue of producing a final ranking where newer tweets must precede older ones with

[1] http://www.lemurproject.org/indri.php

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

[3] http://lucene.apache.org/java/docs/index.html

[4] http://terrier.org/

a simple re-ranking approach. Initially we produced a standard relevancy-based ranking using a standard IR algorithm and then split the retrieved set into two subsets, at the 30^{th} ranked document. The two sets were separately re-ranked based on the timestamp of the tweets, by placing newer tweets higher than older ones. Effectively, we considered the top 30 documents to be *relevant* and re-ranked them based on their timestamp and the rest of the documents, in ranks lower than 30 as *irrelevant*. In order to unite the two time-dependant lists, every tweet in the second set which was newer than the last document in the first list was removed. That was necessary in order to avoid potential conflicts where for example a tweet in rank 30 would be older than tweets further down the final retrieved list.

In order to solve the issue of using global statistics without future knowledge (from tweets that were published after the timestamp of the query), we simply estimated global statistics, such as IDF values only from tweets that were published before the timestamp of the first query in the query set. Clearly, this solution provides only a temporary, less than ideal fix and more appropriate solutions are required, such as dynamically adding documents to the index, either individually or in batches, in a realtime fashion. As we didn't have enough time to explore such solutions, we considered our approach to be good enough for the time being. Additionally, preliminary experiments indicated that the produced lists with and without future knowledge weren't very different, indicating that the limited time resources that were available would be better invested in other areas.

Based on those fundamental changes to the way that Terrier retrieves documents, in order to satisfy the time-based requirements of the Microblog task, we focused on three main research questions in our experiments. First, driven by the innate limit of the number of characters of tweets, we explored whether Query Expansion [1] would positively help retrieval effectiveness. Secondly, we wanted to find out whether query-term proximity models [3, 6] would help and lastly, we explored whether adding the HTML page that a tweet potentially links to, as part of the indexed tweet provides benefits.

The first two solutions have repeatedly shown to help effectiveness in typical ad-hoc retrieval tasks but it remains an open question whether they are also beneficial in this setting, while the last solution is unique to Tweeter. Potential extensions to our approaches could include utilizing field-based weighting models [7], giving different weights to terms appearing in the tweet itself and the linked document and using external sources, such as wikipedia [8] for query expansion.

Although we implemented the facility to remove documents with identical retrieval scores from the final retrieved lists, as a way of removing duplicate or near-duplicate documents, we disabled it for the submitted experiments.

2 Results

In this section we will present our official submissions to the TREC Microblog Track and the results we got. All of our experiments used the Divergence from Randomness (DFR) framework with the PL2 weighting model [2]. Four retrieved lists of documents were submitted to the official TREC:

PL2.NoQe.NoDm Our baseline approach. PL2 weighting model for relevance with no Query Expansion (*NoQe*) and no Dependence Model for term proximity (*NoDm*).

PL2.NoQe.Sd PL2 weighting model for relevance with sequential Divergence from Randomness based dependence model [6].

PL2.NoQe.Sd.Ext PL2 weighting scheme with no query expansion or dependence model but with the linked page as part of the tweet itself (*Ext*).

PL2.Bo1.Sd.Ext PL2 weighting scheme with Bo1 query expansion [1], sequential dependence model and with the linked page as part of the tweet itself.

Table 1 presents the obtained results for all the submitted runs, using all relevant documents. The official track metric is Precision at rank 30 (i.e., P@30) but for completeness reasons, we also present results at ranks $5, 10, 15$ and 20. The best run for each

Table 1: Precision results using all relevant documents. The baseline is a disjunctive run using Lucene.

P@Rank	**Baseline**	**PL2.NoQe.NoDm**	**PL2.NoQe.Sd**	**PL2.NoQe.Sd.Ext**	**PL2.Bo1.Sd.Ext**
P@5	0.2082	0.3633	0.3592	0.0367	0.0449
P@10	0.1612	0.3082	0.3224	0.0612	0.0633
P@15	0.1306	0.2993	0.3048	0.0599	0.0653
P@20	0.1153	0.2878	0.2939	0.0643	0.0653
P@30	0.1007	0.2782	0.2923	0.0633	0.0660

rank is underlined. For comparison reasons, we also present the results from the official track baseline which is a disjunctive run using Lucene.

One of the most notable things that can be observed in the table is the decreased, almost tenfold, effectiveness of the runs that incorporate the linked page into the actual tweet. The results may indicate an issue in the indexing process (e.g., the format in which the linked page was incorporated into the tweet was incorrect for the Terrier indexer) or a very interesting finding where tweets that have incorporated pages get unexpected relevancy scores, due to the significantly increased document length compared with the tweets that do not contain any linked pages. At this stage we are not sure which of the two is the reason, but intent to study the issue closely.

The baseline run using the PL2 weighting model (*PL2.NoQe.NoDm*) achieves a P@30 value of 0.2782, compared to the attained 0.1007 of the baseline run. Using the sequential dependence model increases the precision to 0.2923, offering some evidence that even in this context term proximity models can offer increased effectiveness.

Concerning the usage of query expansion, despite the aforementioned issue with the incorporated linked pages, the run that utilizes it seems to perform better than the one that doesn't (*PL2.NoQe.Sd.Ext* vs. *PL2.NoQe.Sd*, P@30 0.0660 vs. 0.633 respectively), although the differences are very small to reach definite conclusions.

3 Conclusions

This was the first time that the University of Wolverhampton participated in TREC. Due to limited time resources, we decided to only participate in the Microblog Track. In our initial runs, we were interested to find out whether standard IR techniques would be effective in the specific scenario of retrieving Twitter data. Therefore, we experimented with query expansion techniques and proximity dependence models. Both techniques showed some improvement over the simpler baseline, but more experimental results are required in order to reach any final conclusions.

We also experimented with incorporating the linked page to the tweet that was indexed but the approach resulted in significantly decreased retrieval effectiveness. We are currently studying the reasons for the phenomenon, whether it was the result of an error in the indexing process or a natural outcome of the retrieval method.

References

[1] G. Amati. *Probabilistic Models for Information Retrieval based on Divergence from Randomness.* PhD thesis, School of Computing Science, University of Glasgow, 2003.

[2] B. He and I. Ounis. Term frequency normalisation tuning for bm25 and dfr model. In *In Proceedings of ECIR 2005*, pages 200–214. Springer, 2005.

[3] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 472–479, New York, NY, USA, 2005. ACM.

[4] C. Middleton and R. Baeza-yates. A comparison of open source search engines, 2007.

[5] I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras. Research directions in terrier. *Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper*, 2007.

[6] J. Peng, C. Macdonald, B. He, V. Plachouras, and I. Ounis. Incorporating term dependency in the dfr framework. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 843–844, New York, NY, USA, 2007. ACM.

[7] V. Plachouras and I. Ounis. Multinomial randomness models for retrieval with document fields. In *Proceedings of the 29th European conference on IR research*, ECIR'07, pages 28–39, Berlin, Heidelberg, 2007. Springer-Verlag.

[8] W. Zhang and C. T. Yu. Uic at trec 2007 blog track. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 5-9, 2007*, volume Special Publication 500-274. National Institute of Standards and Technology (NIST), 2007.