University of Glasgow (UGLA_D) at TREC Microblogging 2011: Temporal Pseudo-Relevance Feedback in Microblog Retrieval*

Stewart Whiting, Iraklis A. Klampanos, and Joemon M. Jose
{stewh,iraklis,jj}@dcs.gla.ac.uk

School of Computing Science, University of Glasgow,
Scotland, G12 8QQ, UK.

Abstract. This submission represents a first attempt to apply temporal pseudo-relevance feedback for the microblog context. For our submission to the TREC Microblogging 2011 track we perform two approaches, which serve as our initial bases for retrieval. Our first approach uses the retrieval facilities of a standard MySQL server on a heuristically altered tweet collection. Our second approach intends to improve on this initial retrieval through pseudo-relevance feedback, based upon the temporal profiles of n-grams extracted from the top N relevance feedback tweets. A weighted graph is used to model temporal correlation between n-grams, with a PageRank variant employed to combine both pseudo-relevant document term distribution and temporal collection evidence. Preliminary experiments with the TREC Microblogging 2011 Twitter corpus indicate that through parameter optimisation, retrieval effectiveness can be improved.

1 Introduction and Related Work

The significance of time in information production and consumption has been recognised in information retrieval (IR) research. Temporality of topics, relevance composition and term use has been studied in diverse areas of IR and filtering. Topic detection and tracking (TD&T) has provided many methods to analyse time-stamped collections. Yet, despite the success of these algorithms there has been limited work to exploit the collection's temporal properties during retrieval, especially beyond recency ranking. In this paper we propose a model for temporal pseudo-relevance feedback (PRF) based upon temporal and textual evidence combined.

Microblogging has gained popularity as a means of social discussion and commentary of realtime activity. Users are able to publish short text-based snippets (140 characters) which may typically contain linked references such as web pages, other users or subject tags. The most prevalent microblog service today is Twitter[1]. The temporal dimension of information interactions are especially prominent in Twitter [2]. Particularly apparent is the temporal affinity of words and phrases as they rise and fall with topic discussion. In temporal retrieval models there has been work to incorporate the

* This work has been partly supported by EPSRC (EP/H042857/1).

[1] http://www.twitter.com

long-term linear trends of term occurrence in a collection (which we refer to as 'temporal profiles'). Efron [1] used time-series analysis to provide a global term weighting function based upon the prior observations of a term to improve retrieval. Meanwhile, work by Blind [5] concluded that term selection in PRF can be improved by making use of the correlation between the temporal profiles of n-grams obtained from queries and feedback documents.

Query expansion through PRF is often most effective for short and non-specific queries. Query log analysis observes an average query length of 1.64 and 3.08 words for Twitter and web queries respectively [3]. Therefore, in the case of microblog retrieval, the intention of PRF is: firstly, by increasing the number of query terms with further related terms or tags, more tweets relevant to the topic, but perhaps referencing the topic in different language, may be retrieved. Secondly, by expanding the query, tweets which are more descriptive, and so match more query terms, will be higher scored.

In this paper we propose a model to select PRF n-grams based on a variable mix of both temporal and textual (e.g. TF) evidence. Whilst neither temporal or TF evidence alone is consistently able to perform optimally, we demonstrate that combining both evidence sources leads to, on average, better retrieval performance.

2 Approach

To combine both temporal and TF evidence, we propose that selection of n-grams extracted from feedback tweets should be based on the strength of their temporal correlation with other extracted n-grams (i.e. temporal evidence), in combination with their TF within the feedback tweet set (i.e. TF evidence).

We model temporal and TF evidence as a graph. Vertices are n-grams extracted from relevance feedback tweets and edges are directed and weighted according to the temporal correlation between n-grams. N-gram TF in feedback tweets is modelled as vertex priors to form a complete, edge and vertex-weighted graph.

The PageRank (PR) algorithm is effective for discovering nodes with a high relative importance in a network, in our case n-grams that exhibit strong temporal correlation with other n-grams. To combine TF evidence in this estimation of relative importance, PageRank with Priors (PRwP) [4] extends traditional PR by including vertex priors. Priors influence the likelihood of the random walker jumping to a given vertex when teleporting, if the probability of teleporting is > 0. PRwP has a single parameter regulating teleport probability, β. When $\beta = 1$, the random walker will always teleport, so PageRanks will follow the vertex prior distribution, i.e. only TF evidence. Conversely, when $\beta = 0$, the random walker will never teleport and so will move using edge weight probabilities only, i.e. only temporal evidence. With $0 < \beta < 1$, both temporal and TF evidence will be combined.

The temporal association graph is built using only the temporal profile from the start of the collection to the timestamp of the topic (ignoring future evidence). We filter out n-grams with a temporal profile kurtosis[2] < 5 to reduce graph complexity. Low kurtosis n-grams have no significant temporality (i.e., are mostly constant in their use

[2] Kurtosis is a descriptive statistical measure of the 'peakedness' of time-series data.

over time). In the TREC Microblogging 2011 collection, the kurtosis of 'a' is 1.78, whereas for 'superbowl' it is 25.27. Some issues of this with regard to the limited test collection are discussed in Section 4.

N-grams are necessary at the graph modelling stage as many single word terms may be too ambiguous to have a temporal significance. For many single terms, temporal significance is implied by their context (i.e., bigrams). Although other methods exist, we define the temporal correlation function to be the symmetric Pearson correlation between the temporal profiles of the two n-grams, as used in [5].

3 Experiment and Results

3.1 Methodology

Evaluation is performed on the TREC Microblogging 2011 collection, with non-"en" language tweets removed, with test topics MB001 to MB050 from the TREC Microblogging 2011 track. We prepared two separate, initial runs. The results of these procedures were consequently treated as pseudo-relevant documents for the temporal feedback phase. The basic runs were:

1. (*Submitted as simfoll*) Tweets were indexed using a standard MySQL database and its basic full-text search utilities were applied, after acronyms in queries had been expanded. Post-retrieval a couple of heuristics were applied, altering the final ranking: (a) Retweets and similar tweets were removed, as it is understood that in most cases they do not in general add to the informativeness of the original tweet and (b) the score of tweets was boosted in a manner proportional to the number of followers the author had. Strictly speaking, (b) makes use of "future evidence", however, we do not consider this piece of information to be unrealistic given the lifetime and volume of information of services such as Twitter.
2. (*Not submitted - TF*) Tweets were indexed with Lucene without stemming, with stop-words removed. The top 30 tweets with time-stamps prior to the topic time-stamp were retrieved using the built-in vector-space model. The document scoring function was modified, removing document length normalisation and inverse document frequency (IDF) to provide a TF-only retrieval model. This ensured no "future" evidence beyond the query timestamp, in order to simulate real-time retrieval (i.e. Baseline in Table 1(a)).

Given the baselines 1 and 2, above, PRF was applied to the top 20 relevant tweets retrieved. Both PRF-powered runs were submitted as *simfollTP01* and *tfTP01* respectively, each with $\beta = 0.1$. It is worth noting that in both these runs, the final retrieval using the expanded queries was performed using method 2, above.

Temporal profiles for 1- and 2-grams contained within the TREC Microblogging 2011 corpus were mined at 4 hour intervals, over the 16 day duration of the collection, thus leading to a 96 point temporal profile. An interval of 4 hours provides adequate granularity of the temporal variation of the collection.

As we excluded IDF, a traditional PRF approach baseline such as Rocchio's algorithm was inappropriate. We therefore report the TF-only PRF run (i.e. PRF(TF)) and

the Temporal-only PRF run (i.e. PRF(Temporal)). The extreme β setting of 1 and 0, respectively, to achieve this are explained in Section 2.

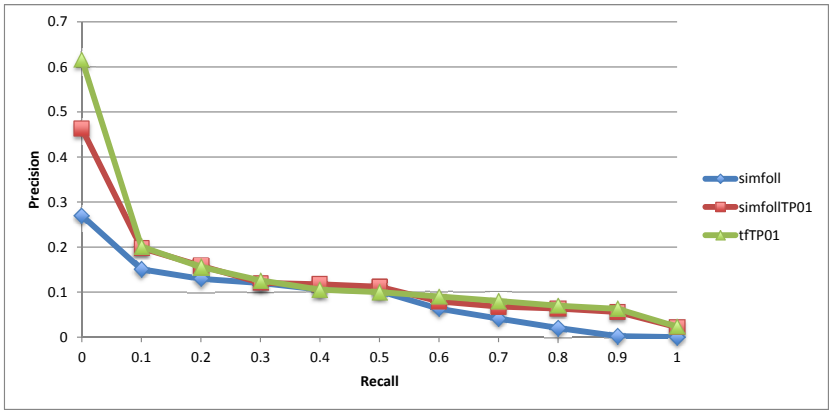
Experimentation with parameter $0 < \beta < 1$ at 0.1 intervals is performed to observe the effect of β on average and per-topic retrieval performance. Table 1(b) analyses theoretical best possible average performance with topic-by-topic β optimisation for each measure (i.e. PRF(TF+Temporal)). Additionally, we report performance using the top 5, 10 or 20 PRF n-grams for query expansion. Original query terms are included with a boost of 1.0, with expansion terms t, boosted at $0.3 \times \frac{t_{pagerank}}{max(pagerank)}$.

We report mean average precision (MAP, with a rank cut-off of 30 tweets) and precision at 5, 10 and 30 (P@5/10/30). While in real-time search, precision models average use patterns accurately, there are other search tasks where recall should also be taken into account, for example, searching for time-insensitive answers in both recent and long-term archives, etc.

3.2 Results and Discussion

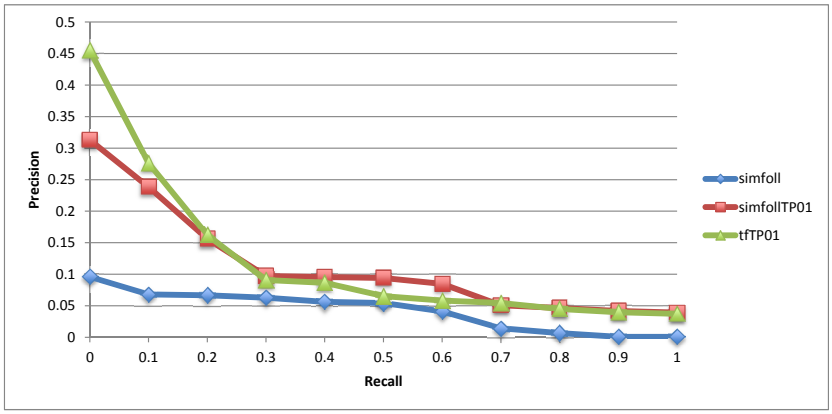
A direct comparison of our three submitted runs is depicted in Figures 1 and 2, where we show how they performed at different recall levels for all relevant and highly relevant tweets respectively. These measures are based on the track evaluation methodology: the 30 most recent tweets retrieved for each topic, regardless of rank. For both all relevant and hi-relevant tweets, tfTP01 appears to outperform the other two runs at all but the highest recall levels.

Fig. 1. Interpolated recall/precision for all relevant tweets, for all 3 submitted runs.



Results not submitted to TREC are reported in Table 1(a). They show that whilst neither PRF(TF) or PRF(Temporal) are able to improve the baseline MAP, PRF(Temporal) is able to outperform both the Baseline and PRF(TF) for all precision measures, albeit marginally for P@5. However, only for P@30 with PRF(Temporal) with 20 expansion

Fig. 2. Interpolated recall/precision for hi-relevant tweets, for all 3 submitted runs.



n-grams is there statistical significance. For this run, P@30 is improved substantially by 16%. Although not reported in this paper, there are $0 < \beta < 1$ settings outperforming the Baseline for all measures.

Without a means of automatically setting β per topic yet, Table 1(b) suggests that with optimisation there is potential for a large improvement on Baseline performance for all measures. Statistical significance for all 5 and 10 expansion n-gram measures indicates that optimising β is possible. Equally apparent is that 5 or 10 n-gram expansion is most effective for all measures. To achieve the highest precision when using less n-grams, relying on temporal evidence is best with a lower β (and β variance) on average. In comparison, for optimal MAP a higher β is necessary on average.

Table 1. Non-submitted runs: reporting MAP, P@5, P@10 and P@30 for Baseline, PRF(TF), PRF(Temporal) and PRF(TF+Temporal). Best performing run for each metric is highlighted. Paired t-test statistical significance is denoted as * being $p < 0.05$.

(a) Baseline, PRF(TF) and PRF(Temporal) results.

Run	MAP	P@5	P@10	P@30
Baseline	**0.1724**	0.4816	0.4286	0.315
PRF(TF):				
5 expansion n-grams	0.1553	*0.4163	0.4184	0.3293
10 expansion n-grams	0.1674	0.4694	0.4306	*0.351
20 expansion n-grams	0.1613	0.4653	0.4306	*0.3483
PRF(Temporal):				
5 expansion n-grams	0.1667	0.4449	0.4347	0.3361
10 expansion n-grams	0.1528	0.4612	0.4224	0.3381
20 expansion n-grams	0.1691	**0.4857**	**0.4469**	***0.3653**

(b) PRF(TF+Temporal) results.

Run	MAP	P@5	P@10	P@30
PRF(TF+Temporal):				
5 expansion n-grams	*0.2177	***0.5837**	***0.5245**	*0.3891
Avg. β	0.32	0.16	0.17	0.17
β Std. Dev.	0.33	0.20	0.19	0.19
10 expansion n-grams	***0.2258**	***0.5837**	*0.5184	***0.4088**
Avg. β	0.38	0.15	0.20	0.19
β Std. Dev.	0.36	0.13	0.22	0.21
20 expansion n-grams	0.2019	0.5429	*0.5143	*0.4048
Avg. β	0.39	0.54	0.39	0.47
β Std. Dev.	0.38	0.46	0.42	0.43

4 Conclusion

Temporal evidence is valuable for PRF in microblogging retrieval scenarios and is more effective than both TF-based PRF and non-PRF retrieval. Even though we only use a rudimentary baseline, state-of-the-art approaches may provide a stronger starting point upon which to maximise the effect of temporal PRF.

In some topics (e.g. MB002: "2022 FIFA soccer") seemingly relevant n-grams are excluded with low temporal profile kurtosis (e.g. '#fifa', '#qatar' and 'world cup'). This is likely due to the limited collection period, as they are unlikely to have a low kurtosis over a longer period. A longer-term collection may achieve better temporal PRF results.

Using TF and temporal evidence together in the oracle optimised run has indicated a strong potential for a per-topic β setting for all performance measures. Future work will concentrate on features to adaptively set this parameter.

References

1. M. Efron. Linear time series models for term weighting in information retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 61:1299–1312, July 2010.
2. M. Efron. Information search and retrieval in microblogs. *J. Am. Soc. Inf. Sci. Technol.*, 62:996–1008, June 2011.
3. J. Teevan, D. Ramage, and M. R. Morris. #twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 35–44, New York, NY, USA, 2011. ACM.
4. S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 266–275, New York, NY, USA, 2003. ACM.
5. S. Whiting, Y. Moshfeghi, and J. M. Jose. Exploring term temporality for pseudo-relevance feedback. In *SIGIR*, pages 1245–1246, 2011.