An Investigation of Term Weighting Approaches for Microblog Retrieval

Paul Ferguson[1], Neil O'Hare[2], James Lanagan[3], Owen Phelan[4], and Kevin McCarthy[4]

[1] Xiam, a Qualcomm Company, Dublin, Ireland. pferguso@qualcomm.com **
[2] Yahoo! Research, Barcelona, Spain. nohare@yahoo-inc.com **
[3] Technicolor R&I, Rennes, France. james.lanagan@technicolor.com **
[4] CLARITY: Centre for Sensor Web Technologies, UCD, Ireland.
{kevin.mccarthy,owen.phelan}@ucd.ie

Abstract. The use of effective term frequency weighting and document length normalisation strategies have been shown over a number of decades to have a significant positive effect for document retrieval. When dealing with much shorter documents, such as those obtained from microblogs, it would seem intuitive that these would have less benefit. In this paper we investigate their effect on microblog retrieval performance using the Tweets2011 collection from the TREC 2011 Microblog Track.

Key words: term weighting, document length normalisation, twitter, microblog retrieval, TREC

1 Introduction

Although there has been a significant amount of research using microblog data for various data mining tasks, there has been relatively little work on microblog retrieval. Massoudi et al [3] propose a language model for microblog retrieval, and incorporate quality metrics and use temporal query expansion to improve performance. Duan et al [1] use a learning to rank approach to combine standard information retrieval ranking with a number of twitter-specific features and authority features. Work such as this acknowledges that standard information retrieval techniques, which use term frequency, document length and inverse document frequency, are unlikely for perform optimally, due to the short document length, but they still use them as a baseline. To the best of our knowledge, no one has yet directly evaluated the applicability of such term weighting approaches to microblog retrieval. Massoudi et al [3] do, in fact, ignore term frequency in their work, but they do not investigate what effect this has on performance, and their language model retrieval framework makes implicit use of document length.

In this work we ask are term frequency and document length normalisation useful for the microblog retrieval task? We examine the influence of term frequency statistics and document length normalisation on microblog retrieval, and

** This work was carried out while Paul Ferguson, Neil O'Hare and James Lanagan were at CLARITY: Centre for Sensor Web Technologies, Dublin City University.

show that not only are these features ineffective for this task, but they actually harm performance. In the next Section we describe the ranking algorithm that we use, and how the parameters of this model can isolate the effects of term frequency and document length normalisation. Our experiments and results are detailed in Section 3, and in Section 4 we draw conclusions from this work.

2 Ranking Microblogs

Standard information retrieval ranking algorithms calculate the relevance of documents with respect to a query based on the frequency of the query terms in a document, normalized by the length of the document, and they use inverse document frequency statistics to downweight non-discrimimnative terms that occur in many documents. The Okapi BM25 model [2] is one such approach, and includes parameters that control the influence of each of these features. It computes the similarity of a document d to query q, containing the terms t, as:

$$bm25(q,d) = \sum_{t \epsilon q} log\left(\frac{N - df_i + 0.5}{df_i + 0.5}\right) \times \frac{(k_1+1)tf_i}{k_1((1-b) + b\frac{dl}{avdl}) + tf_i} \quad (1)$$

Here tf_i represents the document term frequency, dl is the document length and $avdl$ is the average document length in the collection. Adjusting the k_1 and b parameters varies the influence of the term frequency and the document length normalisation. The k_1 parameter controls the influence of term frequency: a value of 0 means that the term frequency is ignored completely, while a higher value increases its' influence. The b parameter adjusts the document length normalisation: b approaching 1 increases the influence of document length normalisation, while a value of 0 results in no document length normalisation. The BM25 model has been extensively tested on the TREC test collections, and has consistently been shown to perform effectively. For this reason, and because it allows for the adjustment of the term frequency weighting and document length normalisation through the k_1 and b parameters, we use this model for ranking microblogs.

When term frequency and document length normalisation are ignored, microblogs containing the same subset of query terms will have an identical score; we re-rank microblogs with tied scores based on recency.

3 Experiments

For our experiments we use the Tweets2011 corpus, developed for the TREC 2011 Microblog Track [4]. The corpus consists of approximately 16 million tweets, gathered over a two week period (January 24th - February 8th, 2011). 50 test topics were developed by the track organizers, each with a timestamp. The real-time search task was to return relevant microblogs from before the query timestamp, ranked in reverse chronological order. The main evaluation measure was Precision@30 (P30). Relevance judgements were created using a pool of 184 official

submitted runs from 59 participating groups. A ternary relevance rating scheme was used, where a tweet could be labelled as *not relevant*, *minimally relevant* or *highly relevant*. Results are reported separately for *all relevant* and for *highly relevant* conditions, with the *all relevant* condition considering both marginally relevant and highly relevant tweets relevant, whereas the *highly relevant* condition only considers highly relevant tweets. Only 33 topics had tweets judged to be highly relevant, so highly relevant results are calculated on this subset of topics. For this evaluation, we rank tweets with using the BM25 model as described above, then take the top 30 most relevant tweets and re-rank them by time.

Figure 1 plots the P30 performance (y-axis) against values for the k_1 (x-axis) and b (z-axis) parameters. These result show a clear upward trend as the value of each of these parameters is decreased, approaching their optimum values as they near 0, i.e. when term frequency and document length are ignored. There is a small improvement in performance at $k_1 = 0.1$ (see Fig. 2), with P30 improving from 0.4211 to 0.4259 for *all relevant*, and from 0.1434 to 0.1495 for *highly relevant*, indicating that any benefit from using term frequency is minimal, and in both cases these increases are not statistically significant (using the t-test).

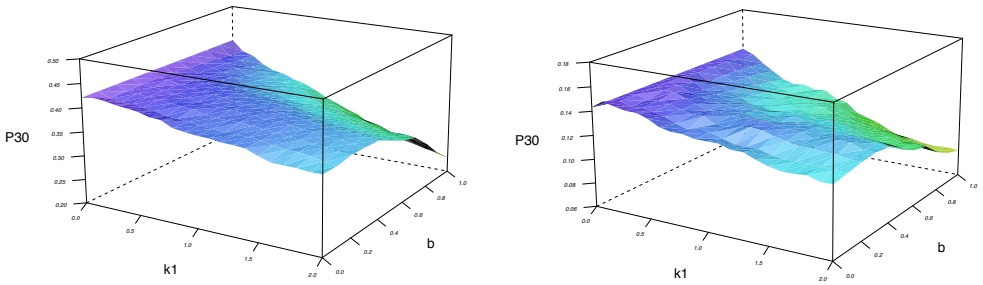

Fig. 1. The effect on P30 as the influence of the term weighting (k1) and document length normalisation (b) are changed.

Document length normalisation always harms performance, which we believe is because it tends to boost the scores of shorter documents, whereas short microblogs are more likely to be noisy and of poor quality. The impact of term frequency and length normal isation can be quite dramatic: typical default BM25 parameters ($k_1 = 1.2$, $b = 0.75$) give P30 of 0.3435 for *all relevant* and 0.1131 for *highly relevant*, which compares with scores of 0.4211 and 0.1434 when both parameters are 0. This improvement over 'default' parameters for *all relevant* was found to be highly statistically significant using a t-test ($p = 0.0013$), while for *highly relevant* the a p value of 0.077 was found.

In our official TREC submissions, the baseline run ($b = k_1 = 0$, *clarity1* in [4]) ranked 5th out of 59 groups for all submissions, and 3rd if only considering runs not using future data (after the query timestamp) or external resources.

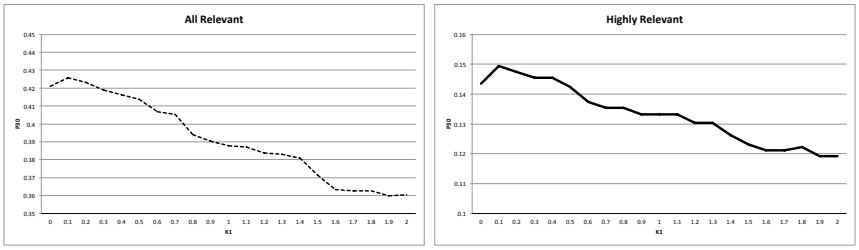

Fig. 2. P30 as the influence of the term frequency (k_1) parameter is changed ($b = 0$).

4 Conclusions

Although it seems intuitive that standard term weighting approaches may not be effective for very short documents, researchers have nevertheless tended to use them as a baseline, which they seek to improve. To our knowledge, this is the first study that examines the applicability of term frequency statistics and document length normalisation to microblog retrieval. The results indicate that document length normalisation always harms performance, and the benefit from incorporating term frequency statistics is minor. The negative influence of document length normalisation also suggests that language model approaches to retrieval will not perform optimally, as they always make implicit use of document length when calculating the probability of a term given a document model. Finally, it would be interesting the examine if this behaviour is replicated in other domains where documents are very short: SMS retrieval, sentence retrieval and multimedia retrieval (where only a few words may describe with each item).

Acknowledgments. This work is supported by Science Foundation Ireland under grant 07/CE/I1147.

References

1. Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum. An Empirical Study on Learning to Rank of Tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 295–303, Beijing, China, August 2010.
2. K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. In *Information Processing and Management*, pages 779–840, 2000.
3. K. Massoudi, E. Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts. In *ECIR 2011: 33rd European Conference on Information Retrieval*, Dublin, 2011. Springer.
4. I. Ounis, J. Lin, and I. Soboroff. Overview of the TREC-2011 Microblog Track. In *TREC 2010 Working Notes*, Gaithersburg, Maryland, USA, 2011. NIST.