

Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts

Kamran Massoudi, Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp

ISLA, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
kamran.massoudi@gmail.com, {e.tsagkias, derijke, w.weerkamp}@uva.nl

Abstract. We propose a retrieval model for searching microblog posts for a given topic of interest. We develop a language modeling approach tailored to microblogging characteristics, where redundancy-based IR methods cannot be used in a straightforward manner. We enhance this model with two groups of quality indicators: textual and microblog specific. Additionally, we propose a dynamic query expansion model for microblog post retrieval. Experimental results on Twitter data reveal the usefulness of boolean search, and demonstrate the utility of quality indicators and query expansion in microblog search.

1 Introduction

Microblogging platforms such as Twitter have become important real-time information resources [4], with a broad range of uses and applications, including event detection [13, 15], and mining consumer and political opinions [6, 14]. In this paper we focus on the task of finding microblog posts for a given query; this task can be thought of as building block of other (mining) tasks that require posts on a specific topic for further downstream processing. The task of searching microblog posts has been studied before: Sakaki et al. [13] investigate the real-time nature of Twitter for event detection and use query expansion to improve recall. Huang et al. [5] analyze so-called hashtag patterns over time and find that hashtags reflect a user’s intent for joining discussions on a topic. TweetMotif [10], an exploratory search application for Twitter, groups messages by significant terms in the query subcorpus.

Items posted on microblogging platforms (like *tweets* and *status updates*) are a special type of user generated content. Their limited size has some interesting effects: (i) people use abbreviations or change spelling to fit their message in the allotted space, giving rise to a rather idiomatic language; (ii) redundancy-based IR methods may not be usable in a straightforward manner to provide effective access to very short documents.

To address the first effect, we introduce a set of *quality indicators* for microblog posts and incorporate these into our retrieval model. In a similar fashion, Weerkamp and de Rijke [16] build on a credibility framework [12] and consider credibility indicators for blog post search. They specifically look at indicators that do not make use of the blogger’s identity, that are textual in nature, and can be reliably estimated using NLP-based methods. Microblogs also have their own characteristics that can be exploited as quality indicators. Kwak et al. [7] study the topological characteristics of microblogs and try to identify influential users in microblogs. Lee et al. [9] extend this by considering the temporal order of information adoption. Recency is considered an important

aspect of microblogs and can be used to improve web ranking for recency sensitive queries [3]. Finally, Cha et al. [2] study the types and degrees of influence within the Twitter network and find that the number of followers represents a user’s popularity, but it is not related to retweets and mentions; the latter two indicate the content value of a microblog post and the value of a user’s name.

To overcome the second effect of redundancy-based IR methods, we re-examine the potential of local *query expansion* for searching microblog posts, using a time-dependent expansion flavor that accounts for the dynamic nature of a topic.

2 Retrieval model and extensions

As a baseline retrieval model, we use a generative language modeling approach (for more details, see for example [1]). An important part of this approach is the estimation of the probability of a term for a given document model (e.g., the microblog post model), $P(t|\theta_d)$. We use Jelinek-Mercer smoothing, and estimate $P(t|\theta_d) = (1 - \lambda) \cdot P(t|d) + \lambda \cdot P(t)$. Due to the short length of microblog posts we expect low language reuse within documents. For this reason, we assume that terms occurring more than once do not add supporting evidence for the relevance of a post d . This directly influences the way we estimate $P(t|\theta_d)$, as detailed in Eq. 1:

$$P(t|d) = \frac{\hat{n}(t,d)}{\sum_{t' \in d} \hat{n}(t',d)} \quad P(t) = \frac{\sum_d \hat{n}(t,d)}{N} \quad \hat{n}(t,d) = \begin{cases} 0 & \text{if } n(t,d) = 0 \\ 1 & \text{if } n(t,d) > 0, \end{cases} \quad (1)$$

where $n(t,d)$ is the term frequency of t in d , and N is the total number of microblog posts (documents) in the collection.

We expect results to contain a fair amount of near-duplicates (so-called reposts in microblogging terminology). Reposts echo another post, but can contain additional information. This extra information can render them more relevant to the query. To deal with this, we proceed as follows. Given a list R of reposts r , originating from post source d , we keep the post r_{best} which has the highest a priori probability according to *recency* and *followers* indicators (see Sec. 2.1).

2.1 Quality indicators

We borrow the following quality indicators for microblog posts from [16]: emoticons, post length, shouting, capitalization, and the existence of hyperlinks. We extend this set of indicators based on the following microblog post characteristics: *reposts*, *followers*, and *recency*, as these represent a certain value of the post.¹ We estimate the repost quality indicator as $P_{repost}(d) = \log(1 + n_{reposts}(d))$. Followers, the number of people “subscribed” to a microblog, are incorporated by putting $P_{followers}(d) = \log(1 + n_{followers}(d))$. Finally, we take recency into account by setting $P_{recency}(d) = e^{-\gamma \cdot (c - c_d)}$, where c is the query time, c_d the post time, and γ the recency parameter.

¹ We write $n_{reposts}(d)$ for the number of reposts of a microblog post d and $n_{followers}(d)$ for the number of followers of the microblog to which post d belongs.

These three microblog specific indicators are combined into a single value, $P_{microblog}$, by taking their average value.

We combine all the indicators to estimate a global credibility prior probability $P(d)$ for a microblog post d , using parameter μ to weigh both indicator groups.

2.2 Query expansion

We capture the dynamics of topics in a microblogging platform using query expansion: After expanding the query, the conditional probability of a term to occur in a query, $P(t|\theta_Q)$, is given by the weighted mixture of the original query Q and the expanded query \hat{Q} , controlled by parameter α . To construct the expanded query, we rank terms according to Eq. 2, and select the top k terms. This model tries to take into account the dynamic nature of microblogging platforms: while a topic evolves, the language usage around it is expected to evolve as well. Consequently, selecting terms temporally closer to query time could result in terms that are more relevant for that point in time. Term scoring becomes a function of time:

$$score(t, Q) = \log \left(\frac{|N_c|}{|\{d : t \in d, d \in N_c\}|} \right) \cdot \sum_{\{d \in N_c : q \in Q \text{ and } t, q \in d\}} e^{-\beta(c-c_d)}, \quad (2)$$

where c is query submission time, c_d is post d 's publication time, N_c is the set of posts that are posted before time c and q . The parameter β controls the contribution of each post to the score of term t based on their posted time. We select only those terms as candidate expansion terms, that occur in more than φ posts.

3 Experimental setup

For the purpose of evaluating microblog post search, we choose to focus on a single target collection in our experimental evaluation, namely Twitter. From the period Nov '09–Apr '10, we drew from Twitter's trending topics, to construct a topic set of 30 queries, with an average length of 1.4 words. We then collected all tweets posted between the last day the topic was "trending" and three days before that day. This resulted in a collection of 110,038,694 tweets. Queries were turned into "topic statements" (in TREC parlance) following the practice of the TREC Blog track [11]. That is, for assessment purposes, each query was equipped with a description of an information need and with a (news) event that corresponds to it. Ground truth was established by pooling the top-20 results of all the retrieval runs that we constructed and assessing the resulting set of posts. Judgments were binary: either relevant or non-relevant.

We do not preprocess the documents (e.g., stopword removal, stemming). Only for the special case of query expansion models, we use a list of 165 English and Spanish stopwords to curate candidate expansion terms. We report on standard IR measures: mean reciprocal rank (MRR), mean average precision (MAP), and precision at position 5 and 10 (P5, P10). Statistical significance is tested using a two-tailed paired t-test and is marked as \blacktriangle (or \blacktriangledown) for significant differences for $\alpha = .01$ and \triangle (and \triangledown) for $\alpha = .05$.

In our language modeling approach, we set the smoothing parameter to $\lambda = 0.15$, reportedly a good value for short queries [17]. For recency, we set $\gamma = 10^{-5}$ assuming c and c_d are measured in seconds. When combining indicators, we set $\mu = 0.375$. For the query expansion model we use the top 10 candidate terms² and set $\alpha = 0.5$, $\beta = 1.5 \cdot 10^{-5}$, and $\varphi = 20$.

4 Results and analysis

The baseline for our retrieval experiments is set to boolean search plus recency (*BS+R*). In this model, relevant posts have to include all query terms and are ranked in reverse chronological order (newer posts rank higher). Our proposed language model is reported as *LM*, the textual quality factor model as *LM-T*, the microblog quality factor model as *LM-M*, and their combination as *LM-T+M*. Results from our retrieval experiments are listed in Table 1. Runs labeled “bl” do not use query expansion and “qe” denotes the use of dynamic query expansion. As a sanity check, we have included a run (“rm2”) that uses an existing standard query expansion model, RM2 [8].

| | <i>BS+R</i> | | <i>LM</i> | | <i>LM-T</i> | | <i>LM-M</i> | | <i>LM-T+M</i> | |
|-----|-------------|---------------------|---------------------|---------------------------|-------------|---------------------------|---------------------|---------------------|---------------------|---------------------------|
| | | bl | rm2 | qe | bl | qe | bl | qe | bl | qe |
| MAP | 0.1640 | 0.0700 [▼] | 0.2390 [△] | 0.4720 [▲] | 0.1930 | 0.4640 [▲] | 0.2190 [▲] | 0.4510 [▲] | 0.2500 [▲] | 0.4820[▲] |
| MRR | 0.7170 | 0.1790 [▼] | 0.5670 [▽] | 0.9670[▲] | 0.7770 | 0.9670[▲] | 0.8220 | 0.9460 [▲] | 0.8230 | 0.9510 [▲] |
| P5 | 0.5870 | 0.1470 [▼] | 0.5530 | 0.9530 [▲] | 0.5800 | 0.9600[▲] | 0.6930 | 0.9330 [▲] | 0.7530 [△] | 0.9530 [▲] |
| P10 | 0.5800 | 0.1570 [▼] | 0.5670 | 0.9600 [▲] | 0.6130 | 0.9630[▲] | 0.6870 [△] | 0.9470 [▲] | 0.7270 [▲] | 0.9570 [▲] |

Table 1: Performance for boolean search plus recency (*BS+R*), language modeling (*LM*), textual (-*T*) and microblog (-*M*) quality factors with dynamic query expansion model enabled (qe) or disabled (bl), with query expansion based on RM2 added for comparison. Significance tested against *BS+R*. Boldface indicates best score in the respective metric.

Our first experiment compares the effectiveness of *BS+R* and *LM*. *BS+R* significantly outperforms *LM*, confirming the usefulness of simple boolean search and recency ranking. Next, we compare our query expansion model to RM. In general, both models show significant improvements in retrieval effectiveness compared to *BS+R* or *LM*. Our dynamic query expansion model outperforms RM2 significantly, and results in a 190% improvement in MAP over *BS+R* and a 500% improvement over *LM*.

Both *LM-T* and *LM-M* improve significantly over *BS+R*, and lead to 180%–260% improvements over *LM*. Their combination, *LM-T+M*, outperforms models based on individual quality indicator sets. Finally, we combine quality indicators and the query expansion model. Expanded *LM-T* and *LM-M*, individually, are unable to outperform query expansion on the baseline *LM*. After combining the indicator sets, however, query expansion achieves the best performance in terms of MAP.

² Based on a small set of preliminary experiments where k varied from 2 to 15 terms, we choose $k = 10$ as a good balance between effectiveness improvement and efficiency.

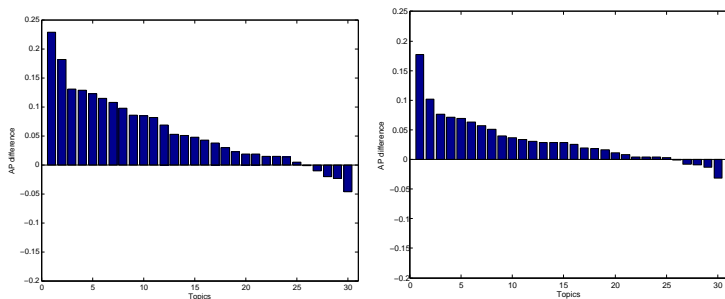


Fig. 1: AP difference between $LM-T+M$ and $LM-T$ (Left) and $LM-M$ (Right).

To gain more insight in the results, we look at per topic differences in average precision (AP). Fig. 1 shows these differences, ordered by improvement in AP, between the combined quality indicators ($LM-T+M$) and the individual sets ($LM-T$ and $LM-M$) without query expansion. The combination outperforms the individual sets on almost all topics. This result strongly suggests that both textual and non-textual indicators should be used in searching microblog posts.

Next, zooming in on query expansion, we see that the query expansion component can help to retrieve real-time tweets. Looking at some examples, we observe that our model brings in top terms like 79 for the query *Woodward*,³ 4 and 4-1 for the query *Messi*.⁴ Top expansion terms are not just single words, but also hyperlinks, hashtags or usernames. For the query *Google Docs*, one of the top expansion terms is <http://bit.ly/4r3sis>, which is the link to Google’s official blog. Examples of hashtags that are returned as top terms are #e09 (elections of 2009) for the query *Chile*, #sotu (State of the Union) for the query *Obama* or #hcr (health care reform) for the query *Health Care*. Finally, @wyclef⁵ is one of the top terms for the query *Earthquake*. The @ appears in all the retweets of user *wyclef*, and as a result tweets from this user will rank higher. The examples above reveal that tokens with numeric or non-alphabetic characters (which are often discarded in traditional retrieval settings) can prove beneficial for query expansion in microblog post search.

5 Conclusions and outlook

We have presented a model for retrieving microblog posts that is enhanced with textual and microblog specific quality indicators and with a dynamic query expansion model. The enhancements were specifically designed to help address the challenges of microblog posts search: rapid language evolution and limited within-document language re-use. We have compared the contribution of our models both individually and com-

³ Actor Edward Woodward passed away at age 79 on the day the query was issued.

⁴ Lionel Messi scored four goals in a match, including his fourth hat-trick of 2010. Barcelona won the match against Arsenal 4-1.

⁵ Musician and Haitian immigrant Wyclef Jean took to his Twitter account to ask fans to help with relief efforts after a 7.0-magnitude quake struck his homeland.

bined on Twitter data, and found that when all models are combined they have a significant positive impact on microblog post retrieval effectiveness. Query expansion on microblog data can be done in a dynamic fashion (taking time into account) and should include specific terms like usernames, hashtags, and links.

In future work, we envisage incorporating quality indicators for reweighing candidate terms for query expansion, exploring diversification techniques for further enhancement of our results and incorporating geolocation features as a prior to users or posts from different locations.

Acknowledgments. This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430 (GALATEAS), by the PROMISE Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 258191, by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.066.512, 612.061.814, 612.061.815, 640.004.802, 380-70-011 and by the Center for Creation, Content and Technology (CCCT).

6 References

- [1] K. Balog, W. Weerkamp, and M. de Rijke. A few examples go a long way: Constructing query models from elaborate query formulations. In *SIGIR '08*, 2008.
- [2] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in Twitter: The million follower fallacy. In *ICWSM '10*, 2010.
- [3] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the essence: improving recency ranking using Twitter data. In *WWW '10*, 2010.
- [4] G. Golovchinsky and M. Efron. Making sense of twitter search. In *CHI '10*, 2010.
- [5] J. Huang, K. M. Thornton, and E. N. Efthimiadis. Conversational tagging in twitter. In *HT '10*, 2010.
- [6] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188, 2009.
- [7] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW '10*, 2010.
- [8] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01*, 2001.
- [9] C. Lee, H. Kwak, H. Park, and S. Moon. Finding influentials based on the temporal order of information adoption in twitter. In *WWW '10*, 2010.
- [10] B. O'Connor, M. Krieger, and D. Ahn. TweetMotif: Exploratory search and topic summarization for Twitter, 2010.
- [11] I. Ounis, C. Macdonald, M. de Rijke, G. Mishne, and I. Soboroff. Overview of the TREC 2006 blog track. In *The Fifteenth Text REtrieval Conference (TREC 2006)*. NIST, 2007.
- [12] V. L. Rubin and E. D. Liddy. Assessing credibility of weblogs. In *AAAI-CAAW '06*, 2006.
- [13] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW '10*, 2010.
- [14] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welp. Predicting elections with twitter: What 140 characters reveal about political sentiment, 2010.
- [15] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *CHI '10*, 2010.
- [16] W. Weerkamp and M. de Rijke. Credibility improves topical blog post retrieval. In *ACL08:HLT*, June 2008.
- [17] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.