York University at TREC 2012: Microblog Track

Zahra Amin Nayeri, Zheng Ye and Jimmy Xiangji Huang

Information Retrieval and Knowledge Management Research Lab
School of Information Technology, York University, Toronto, Canada

Abstract. In this paper, we describe our participation of the ad-hoc task of TREC 2012 Microblog Track. In particular, we evaluate a hybrid retrieval system, which extends the Rocchio's feedback method by incorporating three kinds of IR component techniques. We adapt to the specifics of the microblog search task, giving rise to a highly effective end-to-end search system.

1 Introduction

With the rapid growth of the Internet, the number of short text data on the Web is also growing. There are several models of short text, and microblog is among the most popular ones. In comparison to normal documents. The popularity of Twitter is growing amongst users, as a result ad- hoc information retrieval of this corpus is attracting more researchers. The number of researchers is growing specially after introducing microblog track in TREC 2011. Microblogs nature induces most properties of social networks while lacking some properties of long text documents. One of the vastly noted properties of short text is sparse feature space that makes it difficult to discover correlations among the features. Immediacy and being nonstandard are among the other most important features of short text [9]. Since microblogs are immediate it leads to real-time generation of information and consequently large quantity of the produced short text documents. On the other hand, the content is brief, misspelling is common and non-standard language and structure is frequently used. In TREC 2011 Microblog track 59 different groups of participants submitted 184 runs for a real-time ad hoc search task in Twitter. As the evaluation results show that an appropriate resolution to the real-time search task is yet to be found [14]. Many researchers address tackling problems related to Twitter and propose different approaches for ranking the relevant short text documents. In our participation, we mainly evaluate a recently proposed hybrid retrieval model [8], which has shown to be very effective on a large number of TREC datasets for ad hoc information retrieval.

In particular, this hybrid model extends the Rocchio's feedback method by incorporating three kinds of IR techniques, which are proximity, feedback document quality estimation and query performance prediction techniques, under the pseudo relevance feedback (PRF) framework to boost the overall performance. In our experiments, we test different setting of this hybrid model on the microblog dataset. In the rest of this section, we briefly describe this hybrid model.

2 A Hybrid Retrieval Model

Rocchio's algorithm [5] is a classic framework for implementing (pseudo) relevance feedback via improving the query representation. Although the Rocchio's model has been introduced for many years, it is still effective in obtaining relevant documents. According to [9], "BM25 term weighting coupled with Rocchio feedback remains a strong baseline which is at least as competitive as any language modeling approach for many tasks". However, the traditional Rocchio's model can still be reformed to be better. First, the query term proximity information which has proven to be useful is not considered. Second, Rocchio's algorithm views terms from different feedback documents equally. Intuitively, a candidate expansion term in a document with better quality is more likely to be relevant to the query topic. Third, the interpolation parameter α is always fixed across a group of queries.

In order to address these problems, Ye et al. [8] extend Rocchio's algorithm by refining the query representation as follows.

$$Q_1 = \alpha * (\beta * Q_0 + (1-\beta) * Q_p) + (1-\alpha) * \sum_{r \in R} \frac{r * q(d_r)}{|R|} \quad (1)$$

where β controls how much we rely on the query term proximity information [6], α controls how much we rely on the original query, Q_p is an n-gram of original query terms and $q(d_r)$ is the quality score of document d.

As we can see from Equation 1, this hybrid model is very flexible and can evaluate different techniques. In our experiments, we adopt the co-occurrence interpretation of term proximity to compute Q_p, where the proximity among query terms is represented by the n-gram frequencies and BM25 is used as the weighting model [3]. Full dependencies of query terms are taken into account. For the document quality factor $q(d_r)$, we simple use the normalized scores from the first-pass retrieval for approximation as describe in [7]. For the term weighting formula in the query expansion component, we simply use the Lemur TFIDF formula, which was shown to be surprisingly effective on a number of standard TREC collections in our preliminary experiments.

As we can see from Equation 1, testing different combinations of the component techniques is a straightforward process. In the following, we summarize the component models and the corresponding setting of parameters.

2.1 Parameters

We empirically set parameters as follows; α to 0.6, b in BM25 to 0.3 and β to be 0.2. In our submissions, we did not use the proximity model in run 3, while all the components were used in run 4 with the parameters setting described above.

2.2 submitted Runs

- **YORK1:** We use a weighted Rocchio's feedback model, in which the DFRee weighting model [2] and the KL weighting model (doc=20, term=30, beta=1.4) for query expansion were used.

Table 1. The settings of our submissions

Run	Basic model	Proximity Model	QE Model	Other
YORK1	DFRee	NO	KL weighting Model ($doc = 20,\ term = 30,\ \beta = 1.4$)	NO
YORK2	BM25 ($b = 0.3$)	NO	KL weighting Model ($doc = 20 term = 30 \alpha = 1.4$)	NO
york12mb3	DFRee	NO	KL weighting Model ($doc = 20 term = 30 \alpha = 1$)	After that we conducted filtering according whether the tweet has links and hashtags
york12mb4	DFRee	Yes ($\beta = 0.1, wSize = 8$)	KL weighting Model ($doc = 20 term = 30 \alpha = 1$)	NO

- **YORK2:** We use a weighted Rocchio's feedback model, in which the BM25 (b=0.3) weighting model and the KL weighting model (doc=20 term=30 beta=1.4) for query expansion were used.
- **york12mb3:** We use a weighted Rocchio's feedback model, in which the DFRee weighting model and the KL weighting model (doc=20 term=30 beta=1) for query expansion were used. After that we conducted filtering according whether the tweet has links and hashtags.
- **york12mb4:**
  We use an enhanced Rocchio's feedback model, in which the DFRee weighting model, the proximity model (weight=0.1 + FD + wins=8) and the KL weighting model (doc=20 term=30 beta=1) for query expansion were used.

2.3 A Modified BM25

In order to take into account the specific features of Twitter social network we changed BM25 weighting model for YORK2 run. The new model linearly combines four different scores. As shown in Equation 2.

$$Score(T, D) = w_1 * Score1 + w_2 * Score2 + w_3 * QM_1 + w_4 * QM_2 \quad (2)$$

Where w_i is a real number and;

$$w_2 > w_1 > w_3 > w_4 \quad (3)$$

The first term of Equation uses the traditional BM25 score. BM25 calculates the score as shown in Equation 2.

$$Score1 = \sum_{(q_i \in Q)} \frac{f(q_i, d)}{k_1 * \left((1-b) + b * \frac{|D|}{|avgdl|}\right) + f(q_i, d)} * idf(q_i) \quad (4)$$

The other three terms consider hashtags and links in the tweets. Twitter help center [1] Recommends Twitter users not to use more than two hashtags in each tweet. We investigated hashtags in two cases, first when the hashtag term exists in the topic query, and second for the cases where it does not match query terms. The former is weighted using a similar formulation as BM25 as follows.

if

$$\exists h_i \in Q \tag{5}$$

$$Score2 = \sum_{(q_i \in Q)} \frac{f(h_i, d)}{k_1 * \left((1-b) + b * \frac{|D|}{|avgdl|}\right) + f(h_i, d)} * idf(h_i) \tag{6}$$

For the times that hashtag terms do not occur in the topic we only consider the frequency. This was implemented as shown in Equations 7and 8.

$$QM_1 = logP(h|D) \tag{7}$$

$$P(h|D) = \begin{cases} 1 & n_h > 0 \\ 0 & n_h = 0 \end{cases} \tag{8}$$

Since we did not use any external evidence in our experiments, we did not take the content of URLs into consideration. Equations 9and 10dedicate a score to tweets with URLs.

$$QM_2 = logP(l|D) \tag{9}$$

$$P(l|D) = \begin{cases} 1 & n_l > 0 \\ 0 & n_l = 0 \end{cases} \tag{10}$$

All the coefficients in 2, i.e. w_is, were tuned using microblog track 2011 data set as training data.

3 Experimental Results

In Fig. 1 Average Precision of our four submitted runs on some of the query topics is shown. Different runs show different precision on each topic. In order to investigate the similarity of the four runs and comparing them to median results a mean analysis is performed

As it can be seen from Fig. 2 even though mean average precision of york12mb3 run is better than other three runs, since the confidence intervals in all four runs overlap, we should expect similar results. Mean average precision of the Median results is is less than all of our four runs, but again the overlap indicates that similarity of the results is probable.

Fig. 3 compares the average precision of york12mb3 run to the median of all submitted runs in Microblog track 2012. On most of the topics york12mb3 shows higher precision values. This indicates these results are in top 50% of the overall submitted runs.

[1] https://support.twitter.com/articles/49309-what-are-hashtags-symbols

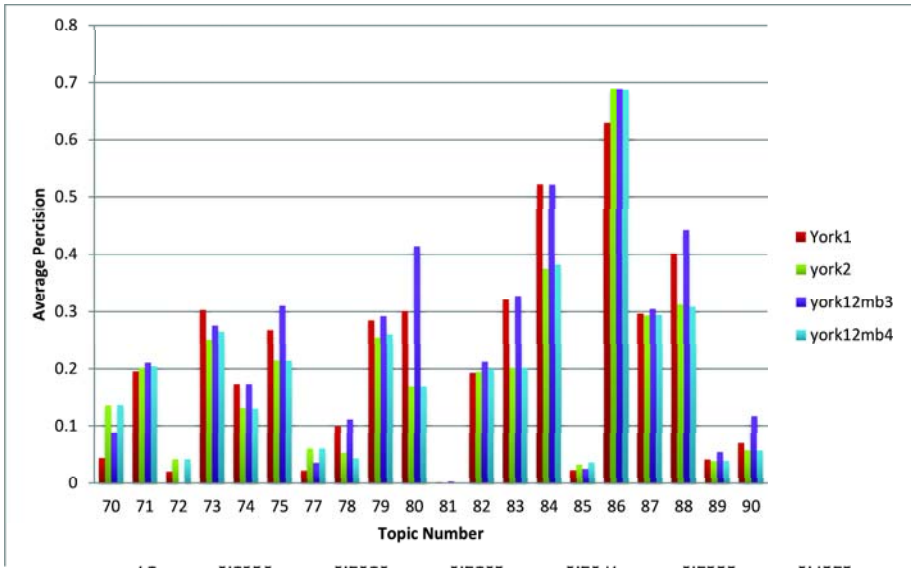

Fig. 1. Illustration of Average Precision for Four Submitted Runs on Topics 70-90

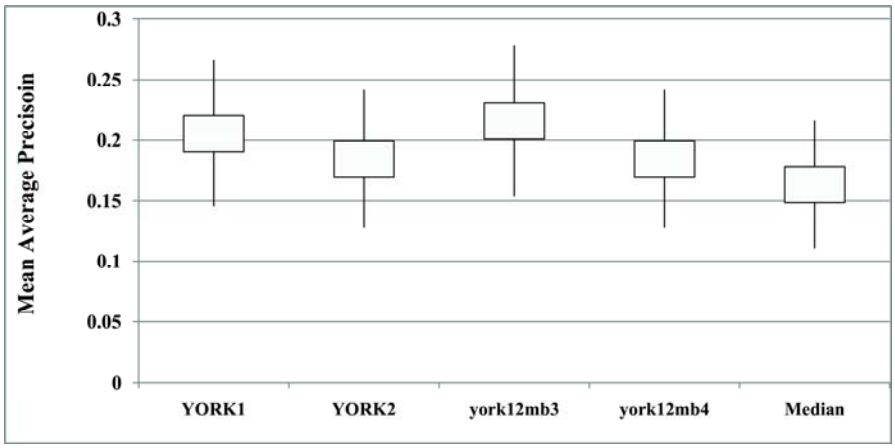

Fig. 2. Mean Comparison of York runs and Median Results

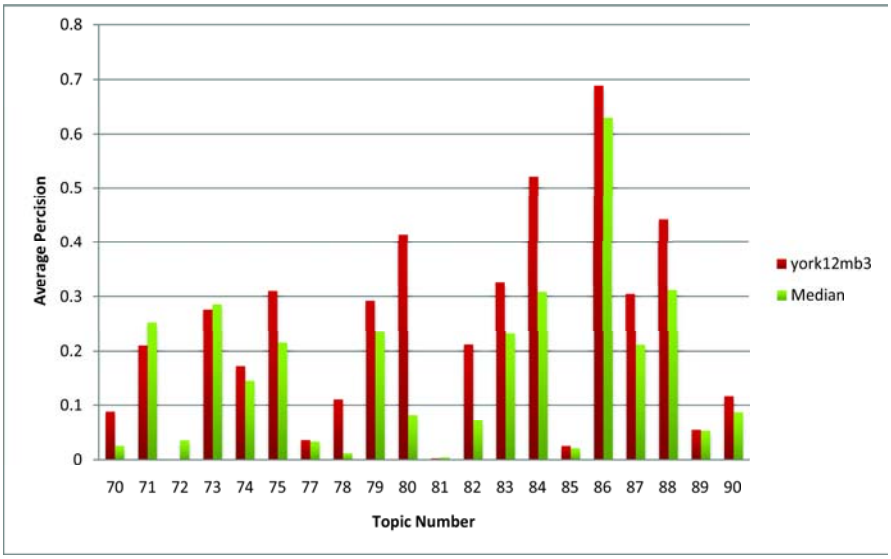

Fig. 3. Comparing Average Precision of york12mb3 to Median Results

4 Conclusion

Our experiments at the TREC 2012 Microblog Track mainly focused on evaluating a recently proposed hybrid retrieval model which has shown to be very effective on a large number of TREC datasets for ad hoc information retrieval. The hybrid model is an extension of Rocchio's feedback method and incorporates three kinds of IR techniques, namely proximity, feedback document quality estimation and query performance prediction techniques, under the pseudo relevance feedback (PRF) framework. In our experiments, we test different settings of this hybrid model on the microblog dataset. In two of the settings we used a modified BM25 specifically tailored for Twitter dataset. Comparing the results to the submitted runs of Microblog track suggests relatively satisfactory results. We plan to make use of external evidence to improve the precision and also extend the weighting model using a larger training set.

Acknowledgements

This research is supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Early Researcher Award/Premier's Research Excellence Award and the IBM Shared University Research (SUR) Award. We also would like to thank IBM Canada for providing IBM BladeCenter blade servers to conduct experiments reported in the paper.

References

1. T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *CIKM*, pages 601–610, 2009.
2. G. Amati and C. J. van Rijsbergen Probabilistic models of information retrieval based on measuring the divergence from randomness. In *ACM Trans. Inf. Syst.*, pages 357–389, 2002.
3. B. He, J. X. Huang, and X. Zhou. Modeling term proximity for probabilistic information retrieval models. *Inf. Sci.*, 181(14):3017–3031, 2011.
4. Y. Lv and C. Zhai. Adaptive relevance feedback in information retrieval. In *CIKM*, pages 255–264, 2009.
5. J. Rocchio. *Relevance feedback in information retrieval*, pages 313–323. Prentice-Hall Englewood Cliffs, 1971.
6. C. J. van Rijsbergen. A theoretical basis for the use of co-occurence data in information retrieval. *Journal of Documentation*, 1977.
7. Z. Ye, B. He, X. Huang, and H. Lin. Revisiting rocchio's relevance feedback algorithm for probabilistic models. In *AAIRS*, pages 151–161, 2010.
8. Z. Ye, X. Huang, and J. Miao. A hybrid model for ad-hoc information retrieval. In *SIGIR*, pages 1025–1026, 2012.
9. C. Zhai. Statistical language models for information retrieval a critical review. *Found. Trends Inf. Retr.*, 2:137–213, 2008.