GUCAS at TREC-2011 Microblog Track

Xin Zhang[1,2], Kai Hui[1,2], Ben He[1,2], and Tiejian Luo[1,2]

[1] Information Dynamic and Engineering Applications Laboratory
[2] Key Laboratory of Computational Geodynamics
Graduate University of Chinese Academy of Sciences
{zhangxin510, huikai10}@mails.gucas.ac.cn, {benhe, tjluo}@gucas.ac.cn

Abstract. The aim of GUCAS's participation in the Microblog track this year is to evaluate the effectiveness of probabilistic retrieval models in combination with various sources of evidence for relevance in the context of the Twitter corpus. In our official runs, we use the PL2F field-based model as the baseline, on top of which query expansion is also applied. In addition, a supplement model combining recency, authority and URL length is developed to retrieve authoritative and timely tweets. Finally, a language filter is used to remove non-English tweets. Our experimental results show that the language filter and URL length filter can benefit the most the retrieval effectiveness. In the following-up experiments, it demonstrates that the results applying the basic models improve siginificantly after removing the retweets in the preliminary results.

1 Introduction

Twitter, as a social network and an information source, is attracting more and more attention [7]. However, Twitter's search engine only provides keyword matching based search results: it presents tweets containing the search query term ranked in chronological order [2]. This mechanism cannot guarantee that the most interesting tweets are top-ranked.

In our experiments, we use the PL2F field-based model for content retrieval to utilize the query term's distribution in the different fractions in the tweets, including the content of the tweets, retweets, and mentions etc. As the content-based PL2F model does not consider the temporal factor during ranking, we develop a supplement model that takes authority, URL length and recency into account. Authority represents the user's influence on others; URL length implicitly represents the richness of the content in the tweet; and recency represents whether the tweet is timely in response to an event, i.e. the query topic.

The rest of the paper is organized as follows. Section 2 introduces the data pre-processing, indexing strategy and the language filter. Sections 3 & 4 introduce the PL2F model and the supplement model combining various sources of evidence. Section 5 presents the experimental results and analysis. Finally, Section 6 concludes our experiments and suggests future research directions.

2 Pre-processing and Indexing

The corpora used in our experiments is in the format of HTML. Before further using it, we first convert the corpora to the TREC format. In particular, in TREC-formatted files, documents are delimited by<DOC></DOC> tags, as in the following example:

```
<DOC>
<DOCNO> 28968126875963392 <DOCNO>
<AUTHOR> TonyFranceOH </AUTHOR>
<TIME> Sun Jan 23 00:11:54 +0000 2011 </TIME>
<AT> </AT>
<BODY> Oh, my GOD </BODY>
<RTAT> </RTAT>
<RT> if today was a boring slop day </RT>
</DOC>
```

In the above example, DOCNO is the tweet id; AUTHOR is the author of the tweet; TIME is the posted time of the tweet; AT contains all the mentioned users in the tweet, except those occurring in RT tweet; RT is the reposted tweet; RTAT indicates the author from which the tweet is retweeted; BODY means the remaining tweet content after removing AT, RTAT, RT.

In our experiments, we build the index using Terrier, version 3.5 [11]. Both direct index and inverted index are built to support retrieval and query expansion. Standard stopword removal and Porter's stemmer are applied.

For the language filter, the LC4j package is used to detect whether a tweet is English or not. It is a language categorization library designed for the Java programming language. It has been designed to be a compact, fast and scalable Java library that implements the algorithms to identify languages using n-grams [12]. In our runs, the detected non-English tweets are removed.

3 Content-based Retrieval

The field-based PL2F model takes the frequency of a query term in different document fields. For instance, it may give a higher weight to a term's appearance in the title than in the body of an HTML document.

PL2F [10] is a per-field derivative of the following PL2 DFR model:

$$\begin{aligned} score(d,Q) = \sum_{t \in Q} \frac{qtw}{tfn+1}(tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn) \cdot \log_2 e + \\ 0.5 \cdot \log_2(2\pi \cdot tfn)) \end{aligned} \tag{1}$$

where $score(d,Q)$ is the relevance score of a document d for a given query Q. λ is the mean and variance of the assumed Poisson distribution of the query term t in the whole collection. qtw is the query term weight, which is given by $qtw = \frac{qtf}{qtf_{max}}$, where qtf is the query term frequency and qtf_{max} is the maximum query term frequency in Q. tfn is a linear combination of the field weight and the normalized term frequency in the i^{th} field as follows:

$$tfn = \sum_{i=1}^{k} w_i \cdot tfn_i \tag{2}$$

where w_i is the weight of the i^{th} field. k is the number of the fields in the document.

The normalized term frequency in the i^{th} field tfn_i is given by Normalization 2:

$$tfn_i = tf_i \cdot \log_2(1 + c_i \frac{avg\_l_i}{l_i}), (c_i > 0) \tag{3}$$

where l_i is and avg_l_i are the document length and the average document length of the i^{th} field respectively. tf_i is the term frequency in the i^{th} field. c_i is a free parameter.

In our experiments, the linear combination weight w_i of the BODY field is set to 7, since it is regarded as the most important part in a tweet. The weight w_i of the other fields, including RT, AT, RTAT, are set to 1. c_i is set to 7 for all the fields.

ignoring the time factor. However, Twitter search requires time together. The supplement model will be discussed in detail in the section 6.

4 Supplement Model

score of a given tweet decreases with its temporal distance to the query. In addition, as authority is shown important in improving query results [1,7],

PL2F involves only the content evidence of the tweet for the relevance ranking. In addition to PL2F, we develop a supplement model that takes authority, URL Length and recency into consideration.

- *Authority* indicates the user's influence on others. We use the number of mentions and the number of retweets in our experiments, because they are only the available evidence in the dataset. The authority score of the tweet's author is computed as follows:

  $$usrScore = (RTAT + 1)^3 \tag{4}$$

  where $RTAT$ is the number of retweets the author has.
  ranks follow to a log-log
  not be influential in terms of spawning retweets or mentions [1], while the number of retweets while mention influence represents the ability of that user to engage others into a conversation. others is also an important metric to estimate the author authority [5]. Finally,

– *URL Length.* As each tweet has the a maximum limit of 140 characters, more detailed information has to be expressed by other Web pages redirected via URLs. Intuitively, a tweet containing URL may convey more information and may be more valuable [3]. In our experiments, we set a Boolean value to indicate whether a URL exists in a given tweet. Moreover, the longer the tweet, the more information the tweet may convey. Therefore, the length of the tweet in characters is also used as an evidence. The URL length score is computed as follows:

$$urlengScore = 0.015 \cdot Length + 0.5 \cdot URL \tag{5}$$

where $Length$ is the number of characters of the tweet. The binary variable URL is 1 if there exists a URL in the tweet, and 0 otherwise.

– *Recency.* One of the goals of the task is real time search. Twitter, as a social network and an information source, can produce a lot of tweets every minute, and it makes real time searching more difficult. Once an event happens, there is a burst time period, during which the event is concerned or mentioned more than other topics. In our experiments, we assign a probability to a tweet, according to the temporal span between the tweet posted time and the query submitted time. Experiments show that an exponential distribution for the tweet prior probability is reasonable. The distribution indicates that tweets with a more recent posted date are assigned higher probabilities. And those tweets that are posted after the query time cannot be retrieval, that is, their probabilities is 0. The recency score is computed as follows:

$$recencyScore = e^{-0.00015.HoursDiff} \tag{6}$$

where $HoursDiff$ is the difference in hours between the tweet's posted time and the query's submitted time.

The final score of a given tweet is computed as follows:

$$finalScore(d, Q) = Score(d, Q)^{1.3} * usrScore * recencyScore * urlengScore \tag{7}$$

Each score is scaled to be within $[0, 1]$ by the following normalized formula:

$$norScore = \frac{(origScore - min) \cdot k_1}{max - min} + k_2 \tag{8}$$

where $origScore$ represents the original score. max and min is the maximum and minimum score of the corresponding evidence. k_1 and k_2 is the empirical parameters. For different sources of evidence, k_1 and k_2 are different. In the next section, we present the experimental results. yet recency seems do the reverse work and it always punishes some highly relevant tweets. for the square is that it can make the PL2F score become the main part and other factors cannot change author and URL&Length result, we multiply them together with powed PL2F score.

5 Experimental Results

We submitted four official runs as follows:

- *IDEABASIC*: A baseline run using PL2F.
- *IDEABASICACT*: A run using supplement model on top of IDEABASIC.
- *IDEABASICQE*: A baseline Run using PL2F with query expansion.
- *IDEAACTQE*: A run using supplement model on top of IDEABASICQE.

In our submitted runs, we rank tweets by their scores given by a combination of PL2F with the supplement model. However, as the official evaluation sorts the tweets by time, we investigate how the evaluation criteria affects the retrieval effectiveness in Table 1.

Table 1. Results obtained by sorting the tweets by scores and by time.

Metrics.	IDEAACTQE	IDEABASIC	IDEABASICACT	IDEABASICQE
P@30ByScore	0.2612	0.2748	0.2701	0.2633
P@30ByTime	0.1177	0.1156	0.1156	0.1177
MAPByScore	0.127	0.1283	0.1312	0.1226
MAPByTime	0.1104	0.1093	0.1093	0.1104

From Table 1, we can see that, our models have relatively weak performance when evaluated under the official setting, while provide better MAP and precision at 30 when the tweets are sorted by their final scores.

improves about 15% and map improves about 2% respectively.

We also conduct additional experiments to examine how the individual evidence affects the retrieval performance. Tables 2 & 3 present the evaluation results without and with query expansion, respectively. From these two tables, we can see that, the language filter and URL length can benefit the most the retrieval effectiveness. However, authority and recency do not improve the results as expected.

Also in the following-up experiments, after removing retweets, the results using the basic retrieval models improve significantly. For the details, we can refer to Table 4.

Table 2. Evaluation of different sources of evidence on top of baseline IDEABASIC without query expansion.

	Baseline	+recency	+language	+URL length	+authority	all Filters
MAP	0.1342	0.1013, -24.51%	0.129, -3.87%	0.1364, +1.64%	0.1343, +0.07%	0.1312, -2.24%
P@30	0.2728	0.1245, -54.36%	0.2769, +1.50%	0.2803, +2.75%	0.2735, +0.26%	0.2701, -0.99%

Table 3. Evaluation of different sources of evidence on top of baseline IDEABASIC with query expansion.

	Baseline	+recency	+language	+URL length	+authority	all Filters
MAP	0.1335	0.1013, -24.12%	0.1246, -6.67%	0.1381, +3.45%	0.1382, +3.52%	0.1270, -4.87%
P@30	0.2653	0.1347, -49.23%	0.2667, +0.53%	0.2776, +4.64%	0.2639, -0.53%	0.2612, -1.55%

Table 4. Results obtained after removing the retweets.

Metrics.	BM25	PL2	DirichletKL	DFRee
P@30	0.3252	0.3469	0.3449	0.3592

6 Conclusions and Future Work

We have developed a supplement model on top of the field-based PL2F model to combine various sources of evidence, including recency, URL Length, and authority. As shown by the official runs and our preliminary results, apart from URL length, the additional evidence does not improve the retrieval effectiveness as expected. However, the following-up experimental results using the basic retrieval models improve siginificantly after removing the retweets and the effectiveness using the DFRee model can achieve to 0.3592. We are now under an investigation in the reasons for the failure of the use of authority and recency. In future, we plan to propose a new method to incorpurate all the evidence mentioned above.

Acknowledgements

This work is funded in part by Project 61103131/F020511 supported by NSFC, and the President Fund of GUCAS (Y15101FY00).

References

1. M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. *Measuring User Influence in Twitter: The Million Follower Fallacy.* In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM), 2010.
2. 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, COLING '10, pages 295-303, Stroudsburg, PA, USA, 2010.
3. R. Nagmoti, A. Teredesai, and M. De Cock. *Ranking approaches for microblog search.* In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01, WI-IAT '10, pages 153-157, Washington, DC, USA, 2010.
4. A. Pal and S. Counts. *Identifying topical authorities in microblogs.* In Proceedings of the fourth ACM international conference on Web search and data mining, WSDM'11, pages 45-54, New York, NY, USA, 2011.

5. Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, Hongyuan Zha. *Time is of Essence: Improving rency ranking using Twitter data.* In Proceedings of the 19th international conference on World wide web, pages 331-340, Raleigh, North Carolina USA,2010.
6. Miles Efron and Gene Golovchinsky. *Estimation methods for ranking recent information* , In Proceedings of the 34th Annual ACM Conference, pages 495-504, Beijing, China, 2011.
7. H. Kwak, C. Lee, H. Park, and S. Moon. *What is twitter, a social network or a news media?* . In Proceedings of the 19th international conference on World wide web, WWW'10, pages 591-600, New York, NY, USA, ACM, 2010.
8. 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,ACM, 2011.
9. X. Li, W. Bruce Croft *Time-based language models* . In Proceedings of the twelfth international conference on Information and knowledge management, 2003.
10. C. Macdonald, V. Plachouras, B. He, C. Lioma, and I. Ounis. *University of Glasgow at WebCLEF 2005: Experiments in Per-Field Normalization and Language Specific Stemming* . In Proceedings of CLEF 2005.
11. Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma. *Terrier: A High Performance and Scalable Information Retrieval Platform.* In Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR), 10th August, 2006. Seattle, Washington, USA, 2006.
12. `http://olivo.net/software/lc4j/`